Machine Learning for Asthma Management: An Integrated Framework Combining Environmental Prediction of Respiratory Hospitalizations with Digital Biomarkers of Adherence to Diaphragmatic Breathing

Daniel Pereira Ferreira *, Gabriel Fuscald Scursone, Diana Francisca Adamatti

Posted Date: 16 June 2026

doi: 10.20944/preprints202606.1217.v1

Keywords: asthma; machine learning; air pollution; meteorology; CatBoost; XGBoost; SHAP; digital biomarkers; diaphragmatic breathing; environmental health

 doi:10.20944/preprints202606.1217.v1

Article

Machine Learning for Asthma Management: An Integrated Framework Combining Environmental Prediction of Respiratory Hospitalizations with Digital Biomarkers of Adherence to Diaphragmatic Breathing

Daniel Pereira Ferreira *, Gabriel Fuscald Scursone and Diana Francisca Adamatti

Graduate Program in Computational Modeling, Federal University of Rio Grande, Rio Grande, RS, Brazil

* Correspondence: daniielferr@hotmail.com

Abstract

(1) Background: Asthma is a chronic respiratory disease shaped by environmental, meteorological, and behavioral factors. Although the literature has advanced in predicting respiratory outcomes and in deploying digital technologies for therapeutic support, few approaches integrate population surveillance and individual monitoring within the same analytical framework. **(2) Methods**: This work developed an integrated machine learning framework composed of two complementary studies. Study 1 modeled the daily count of hospital admissions for all respiratory diseases (chapter X of the ICD-10, which includes asthma, COPD, and respiratory tract infections), denoted HOSPCIDX, in the municipality of São Paulo (Brazil). It drew on six years of data (2017 to 2022) from SIH/SUS via PCDaS/Fiocruz, CETESB, and INMET, and applied a hybrid architecture combining ElasticNet, residual CatBoost, direct CatBoost, and adaptive blending, validated through walk-forward over 30 bimonthly folds across the 2018 to 2022 period (2017 was reserved for lag construction). Study 2 focused specifically on pediatric asthma and analyzed 913 qualified records from the Respire Bem system, collected from patients aged 6 to 16 years, using XGBoost and Random Forest models. The clinical outcomes (Asthma Control Test and salivary cortisol) were generated by evidence-based synthetic simulation. **(3) Results**: In Study 1, the hybrid model achieved a mean MAE of 18.22, a mean RMSE of 23.99, a mean R^2 of 0.675, and a mean skill gain of 41.5% over the seasonal baseline. SHAP analysis identified mean temperature, PM2.5, NO2, and CO as the main predictive drivers of respiratory hospitalizations. In Study 2, XGBoost reached an R^2 of 0.80 for the simulated Asthma Control Test and 0.78 for simulated salivary cortisol, with self-reported sentiment emerging as the leading digital biomarker. **(4) Conclusions**: The proposed framework demonstrates the feasibility of a dual analytical architecture for asthma management, combining environmental prediction at the population level with digital monitoring at the individual level. Study 1 provides robust predictive validation with real data, while Study 2 represents an exploratory stage based on real behavioral data and simulated clinical outcomes, which calls for prospective validation with direct clinical and biological measurements.

Keywords: asthma; machine learning; air pollution; meteorology; CatBoost; XGBoost; SHAP; digital biomarkers; diaphragmatic breathing; environmental health

1. Introduction

1.1. Global and Brazilian Burden of Asthma

Asthma is one of the most prevalent chronic respiratory diseases worldwide, affecting approximately 262 million people and causing about 455,000 deaths in 2019 according to estimates

 doi:10.20944/preprints202606.1217.v1

by the World Health Organization [1]. From a broader perspective, the State of Global Air 2024 report attributes 8.1 million deaths in 2021 to air pollution, with a strong contribution from respiratory diseases and from the worsening of asthma in pediatric and elderly populations [2]. In Brazil, the disease accounts for a substantial share of avoidable pediatric hospitalizations, generating direct costs for the Unified Health System (SUS) and indirect costs associated with school and work absenteeism, particularly in large metropolitan centers [3].

In the municipality of São Paulo, the largest urban center in the Southern Hemisphere, the coexistence of high concentrations of atmospheric pollutants with marked meteorological variations amplifies population exposure to respiratory risk factors. The intense vehicle fleet, the industrial complexity, the high population density, and the consolidated environmental and meteorological monitoring networks make the city an urban laboratory of international relevance for research in environmental health. In addition, the Unified Health System provides, through DATASUS, national records of hospital admissions with wide coverage and standardized organization by the International Classification of Diseases, 10th revision (ICD-10), which allows the construction of robust time series for predictive modeling.

Despite the therapeutic advances of the last decades, asthma control remains suboptimal for a substantial proportion of patients, a phenomenon often attributed to low treatment adherence, persistent exposure to environmental triggers, and the limited capacity of health systems to anticipate exacerbations. The combination of these factors motivates the search for management strategies that articulate population surveillance and individual intervention, ideally supported by modeling tools able to process heterogeneous data sources [4,5]. International guidelines, notably the Global Initiative for Asthma (GINA) strategy [6], reinforce individualized care plans, patient education, and trigger avoidance as complementary pillars to pharmacotherapy.

1.2. Environmental and Meteorological Determinants of Respiratory Hospitalizations

The relationship between air pollution and respiratory outcomes is widely documented in the epidemiological literature. Fine particulate matter (PM2.5), inhalable particulate matter (PM10), nitrogen dioxide (NO2), tropospheric ozone (O3), sulfur dioxide (SO2), and carbon monoxide (CO) are, individually or jointly, associated with increased hospital admissions for asthma, chronic obstructive pulmonary disease, and respiratory tract infections [5,7]. In a comprehensive study of 652 cities, Liu et al. [8] showed that increases in PM2.5 are consistently associated with higher daily mortality from respiratory causes, with larger magnitudes in densely populated urban areas.

Meteorological variables modulate these associations through multiple mechanisms, such as the dispersion and deposition of pollutants, changes in the dynamics of aeroallergens, and thermal stress on vulnerable populations [9,10]. Achebak et al. [11], in a multicenter European study, showed that variations in ambient temperature explain a significant share of the interannual variability of mortality from respiratory causes, with a particularly strong effect of cold. Romaszko-Wojtowicz et al. [12] documented a consistent association between seasonal biometeorological conditions, particulate matter concentrations, and hospital admissions for asthma and COPD across a ten-year time series in northeastern Europe.

Classical studies in the time series rely on distributed lag non-linear models (DLNM), typically implemented within generalized linear or generalized additive frameworks [13]. These approaches, although well grounded, may underestimate nonlinearities, interactions between variables, and regime shifts, limitations that can be overcome by machine learning methods, especially those based on tree ensemble algorithms [14,15].

1.3. Non-Pharmacological Treatment and Digital Biomarkers

Therapeutic adherence is a critical determinant of asthma control, particularly in pediatric populations, where reported non-adherence rates range from 30% to 70% in different settings. This high variability reflects difficulties with sustained engagement, parental supervision, and the need

 doi:10.20944/preprints202606.1217.v1

for behavioral reinforcement strategies, which positions adherence support as a priority area for digital health interventions.

Beyond pharmacotherapy, non-pharmacological techniques have gained attention as adjuvants in asthma management. Diaphragmatic breathing stands out for its simplicity, low cost, and potential to reduce stress, which is often associated with the intensification of asthma crises. This effect has been shown in clinical trials by improvements in the Asthma Control Test and by reductions in salivary cortisol, a validated biomarker of physiological stress [16]. Despite this clinical evidence, transposing the technique to the patient's daily routine remains challenging, especially in pediatric populations, where adherence depends on sustained engagement and on technological mediators.

The Respire Bem system [17] was designed as a digital interface to support the daily practice of diaphragmatic breathing, recording adherence variables (frequency, completeness, and duration of breathing cycles) and the subjective state self-reported by the patient. The systematic analysis of these records makes it possible to operationalize digital biomarkers, defined as computational proxies of validated clinical outcomes. This approach is aligned with contemporary trends in personalized medicine and digital health. The systematic review by Ferreira et al. [18] synthesized the recent literature on machine learning applied to asthma and identified a specific gap in predictive models focused on monitoring adherence and the efficacy of complementary therapies through digital health interventions, a gap that the present work helps to fill.

1.4. Machine Learning Applied to Respiratory Health

Machine learning (ML) methods have emerged as competitive tools for prediction in health, particularly in epidemiological time series and longitudinal individual data. Gradient boosting algorithms, such as CatBoost [19] and XGBoost [20], combine high predictive capacity, robustness to missing data, and interpretability through techniques such as SHAP (Shapley Additive exPlanations) [21]. In parallel, temporal validation schemes, especially walk-forward, ensure that the performance estimate reflects the prospective use of the model and avoids leakage of future information into the training set [22,23].

Recent studies corroborate the applicability of these methods to the specific problem of predicting respiratory hospitalizations. Cabral-Miranda et al. [24] presented an artificial intelligence platform for predicting pediatric respiratory care based on clinical, environmental, and climatic factors, with competitive performance in a Brazilian setting. Barnett-Itzhaki et al. [25] showed the usefulness of ML models for predicting pediatric hospitalizations as a function of air pollution and humidity. Monteiro Martins et al. [26], in a recent systematic review, synthesized forty studies that use environmental predictors to forecast hospital visits and admissions and identified relevant gaps, especially the scarcity of validation and interpretability analyses that engage with the epidemiological literature.

1.5. Identified Gap and Contributions

Although the volume of studies on the topic has grown, many of them still analyze these aspects in isolation. On one side, there is research on the environmental impacts on respiratory health. On the other, there are studies on technology-mediated behavioral interventions. Few works integrate these dimensions within a single analysis. This integration matters because such factors share the same patient trajectory: environmental exposure can trigger crises, the health system absorbs their impacts, and individual interventions can reduce the frequency and severity of these episodes. The present work offers four main contributions:

- It presents a hybrid machine learning architecture composed of ElasticNet, residual CatBoost, direct CatBoost, and adaptive blending, validated by walk-forward over 30 bimonthly folds, covering six years of data (2017 to 2022) with a five-year prospective test window and including the COVID-19 pandemic shock;

 doi:10.20944/preprints202606.1217.v1

- It quantifies, through SHAP values, the relative contribution and the direction of the effect of pollutants and meteorological variables, identifying mean temperature, PM2.5, NO2, and CO as the main factors associated with respiratory hospital admissions;
- It integrates digital biomarkers derived from adherence to diaphragmatic breathing, captured by the Respire Bem system, articulating the population-level environmental axis with the individual behavioral axis;
- It provides a reproducible computational pipeline, with fixed random seed and standardized audit artifacts, in agreement with the principles of reproducible research in data science applied to public health.

1.6. Scientific Hypothesis and Objectives

We hypothesize that (a) a hybrid machine learning architecture combining environmental, meteorological, and temporal features outperforms classical baselines (weekly, seasonal, and ElasticNet) in predicting daily respiratory hospitalizations, with a skill gain of at least 30% in RMSE over the seasonal baseline, and (b) digital biomarkers extracted from a non-pharmacological asthma intervention system can serve as computational proxies of the expected clinical response, when calibrated against controlled clinical evidence. The integration of these two complementary layers, environmental at the population scale and behavioral at the individual scale, is expected to support a feasible framework for combined surveillance and clinical management of asthma.

The general objective is to develop and validate an integrated machine learning framework that connects the population level, represented by the daily HOSPCIDX count, with the individual level, represented by the digital biomarkers of the Respire Bem system. The specific objectives are (i) to build a unified daily frequency database integrating records from SIH/SUS via PCDaS/Fiocruz, CETESB, and INMET, with deterministic linkage for the municipality of São Paulo (Brazil) in the 2017 to 2022 period; (ii) to implement a hybrid predictive pipeline with walk-forward validation over 30 bimonthly folds; (iii) to characterize the importance and the direction of environmental drivers through SHAP values restricted to raw variables; (iv) to assess the robustness of performance under the COVID-19 pandemic shock; and (v) to validate the feasibility of a two-layer analytical structure for asthma management, through the joint use of the population-level environmental model and the individual digital biomarker models.

The remainder of the article follows the structure: Section 2 details the methodological design of the two studies. Section 3 presents the results, including the sensitivity analysis. Section 4 discusses the findings in dialogue with the recent literature. Section 5 concludes by synthesizing contributions, limitations, and future perspectives.

2. Materials and Methods

2.1. Overview of the Integrated Design

This work adopts an integrated design composed of two complementary studies. Study 1 is an ecological time series study with daily resolution, a predictive approach, and a focus on environmental health. It addresses the relationship between environmental exposures, meteorological variables, and the daily count of respiratory hospitalizations (HOSPCIDX). Study 2 has a longitudinal individual nature and investigates the relationship between adherence metrics to diaphragmatic breathing, mediated by the Respire Bem system, and clinical outcomes simulated from evidence reported in a controlled clinical trial. The integration between the two studies takes place in the discussion phase, where the articulation between the models is proposed as a dual system of surveillance and intervention for asthma.

2.2. Study 1: Environmental and Meteorological Modeling of Hospitalizations

2.2.1. Design and Spatiotemporal Scope

 doi:10.20944/preprints202606.1217.v1

The spatial scope corresponds to the municipality of São Paulo, and the temporal scope covers six full years, from January 1, 2017 to December 31, 2022, totaling 2,191 consecutive days. The design is longitudinal and retrospective, since it is based on secondary data already recorded in official information systems, and predictive, since it seeks to estimate the daily number of respiratory admissions from environmental exposures observed on the same day or on previous days. The choice of an ecological design is justified by the collective nature of the research question and is widely used in environmental epidemiology to estimate the aggregate effect of diffuse exposures, such as air pollution and meteorological variations, on hospital outcomes measured at the population scale [27].

2.2.2. Data Sources and Linkage Procedure

Health data were obtained from the Hospital Information System of the Unified Health System (SIH/SUS), administered by the Department of Informatics of the SUS (DATASUS), through the institutional repository of the Data Science Platform Applied to Health of the Oswaldo Cruz Foundation (PCDaS/Fiocruz), which integrates, in a structured way, information on hospital morbidity and mortality across the national territory. The stratification of admissions was conducted by ICD-10 chapter, based on the codes officially provided by DATASUS, ensuring standardization and comparability throughout the study period.

Air pollution data were extracted from the monitoring networks operated by the Environmental Company of the State of São Paulo (CETESB), whose grid of fixed stations distributed across the city forms the largest continuous air quality monitoring network in operation in Brazil. Meteorological data were obtained from the National Institute of Meteorology (INMET), which is responsible for the official network of conventional and automatic meteorological stations in the country.

Integration between the three databases was conducted through a deterministic linkage procedure, that is, direct linkage between records from different sources, based on the exact correspondence of date (daily frequency) and location (municipality of São Paulo). Since the three sources use compatible calendars and temporal granularities, the linkage procedure did not require temporal aggregations or probabilistic keys, which simplified the operation and reduced the risk of biases from granularity mismatch. This procedure ensured the spatiotemporal consistency of the data and supported the construction of the unified daily frequency database used in all subsequent analyses.

2.2.3. Response Variable and Predictor Variables

The response variable is the daily count of hospital admissions for all diseases of the respiratory system (chapter X of the ICD-10, which includes asthma, J45-J46; chronic obstructive pulmonary disease, J44; pneumonia and influenza, J09-J18; and other respiratory tract infections) in the municipality of São Paulo. This variable is denoted HOSPCIDX and is treated as an integer-valued time series, without logarithmic transformation or variance stabilization, in order to preserve the direct interpretation of results in absolute number of admissions. The use of all respiratory admissions, rather than asthma-specific admissions only, was a methodological choice that prioritizes statistical power and the operational scope of population-level surveillance. The implications of this choice for clinical specificity are discussed in Section 4.7 (Limitations).

The predictor variables were organized in two complementary groups, in daily tabular format. The first group comprises six atmospheric pollutants widely recognized in the epidemiological literature as agents directly associated with the worsening of respiratory diseases [5]: inhalable particulate matter (PM10) and fine particulate matter (PM2.5), expressed in $\mu g/m^3$; tropospheric ozone (O3), nitrogen dioxide (NO2), and sulfur dioxide (SO2), expressed in $\mu g/m^3$; and carbon monoxide (CO), expressed in ppm. The second group comprises six meteorological variables, identified as TEMPMED (daily mean temperature, in °C), UMIDRELAT (relative humidity, in %), PRECIPIT (accumulated daily precipitation, in mm), RADIACAO (daily global solar radiation, in kJ/m^2), VELVENTO (mean wind speed, in m/s), and PRESSAOATM (mean atmospheric pressure, in hPa). The full description of the variables used in the study is summarized in Table 1.

 doi:10.20944/preprints202606.1217.v1

Table 1. Description of the variables used in Study 1 for the municipality of São Paulo, 2017 to 2022.

Variable	Type	Unit	Description	Source
PM10	Real	µg/m³	Inhalable particulate matter	CETESB
PM2.5	Real	µg/m³	Fine particulate matter	CETESB
O3	Real	µg/m³	Tropospheric ozone	CETESB
NO2	Real	µg/m³	Nitrogen dioxide	CETESB
SO2	Real	µg/m³	Sulfur dioxide	CETESB
CO	Real	ppm	Carbon monoxide	CETESB
TEMPMED	Real	°C	Daily mean temperature	INMET
UMIDRELAT	Real	%	Relative humidity	INMET
PRECIPIT	Real	mm	Accumulated precipitation	INMET
RADIACAO	Real	kJ/m²	Global solar radiation	INMET
VELVENTO	Real	m/s	Mean wind speed	INMET
PRESSAOATM	Real	hPa	Mean atmospheric pressure	INMET
HOSPCIDX	Integer	admissions	Daily admissions for respiratory diseases (ICD-10, chapter X)	SIH/SUS via PCDaS/Fiocruz

Source: prepared by the authors.

The period covers heterogeneous epidemiological scenarios. The pre-pandemic period (2017 to February 2020) shows relatively stable seasonal patterns in respiratory hospital demand. The critical phase of the COVID-19 pandemic (from March 2020 onwards) brought strong changes in both hospital admissions and atmospheric pollution levels as a consequence of social distancing measures. The recovery of urban activities took place in 2021 and 2022. To allow the model to represent this structural change without being distorted by it, a binary pandemic indicator was incorporated into the feature set, activated from March 1, 2020 and held constant from that date onwards.

2.2.4. Preprocessing and Feature Engineering

Preprocessing comprised four steps: (i) coercion of all columns into standardized numerical types, with non-convertible values flagged as missing; (ii) reindexing of the time series to strict daily frequency; (iii) harmonization of variable names across alternative spellings (for example, PM2.5, PM2,5, and PM_2_5 mapped to a single canonical identifier); and (iv) creation of binary missing-value flags for every variable, preserving information about gaps in monitoring.

A causal protocol for handling missing values was designed as a safety mechanism of the pipeline, ensuring that the imputation of any missing day would use only information observed before it. This choice is consistent with the realistic operational scenario of using the model for public health surveillance. The protocol operates in two sequential steps: a forward fill limited to seven days, which propagates the last observed value to the immediately following days, followed by a causal 21-day rolling median that fills longer gaps using only values prior to the day in question, which avoids leakage of future information. The target variable HOSPCIDX is never imputed, and any day with a missing target is deliberately excluded from the training, validation, and testing stages. In the consolidated database used in the present analysis (see Section 3.1.1), the thirteen variables showed full completeness after deterministic linkage, so the imputation protocol was implemented as a safety net rather than as an extensively triggered step.

Feature engineering covered three groups. The first group includes attributes derived from the target series itself, with lags at 1, 2, 3, 4, 5, 6, 7, 10, 14, 21, 28, 35, 42, 56, 84, 364, and 365 days; rolling means at 7, 14, and 28 days; exponential rolling means with spans of 7, 14, and 28; a seasonal baseline computed as the mean of lags 364 and 365; and differences between lags (lag1 minus lag7, lag7 minus lag14). The second group includes contextual and seasonal attributes, with day of year, day of week, month, week of year, cyclic indicators by sine and cosine, binary indicators for the beginning and end of the month, and a weekend indicator, along with the binary flag for the pandemic period. The third

 doi:10.20944/preprints202606.1217.v1

group consists of attributes derived from environmental exposures, with short lags from 1 to 28 days, differences between rolling means (7 and 14 day anomalies), binary interactions between pollutants and meteorological variables (for example, PM2.5 × relative humidity, O3 × temperature, NO2 × temperature), and indicators of intraday variation.

2.2.5. Hybrid Model Architecture

The proposed model articulates three components within a hybrid architecture with adaptive selection by internal validation. The first component is a regularized linear regressor (ElasticNet) with α = 0.0045 and l1_ratio = 0.08, called the base model, trained only with attributes derived from the target series and from the temporal context. This component captures the linear structure, the regular seasonality, and the trend of the series [28]. The second component is a CatBoost regressor (depth 6, learning rate 0.030, l2_leaf_reg = 3.0, min_data_in_leaf = 40, bagging_temperature = 1.0) trained to predict the residuals of the base model from the exogenous attributes (pollutants and meteorology). CatBoost is able to capture nonlinear relations, threshold effects, and interactions between environmental exposures and the temporal context [19]. The third component is a direct CatBoost regressor, trained on the full set of attributes to predict the target variable in an end-to-end fashion.

For each validation fold, four predictive strategies are evaluated. The first, base only, uses the ElasticNet prediction alone. The second, residual only, sums the ElasticNet prediction with the CatBoost prediction of residuals. The third, direct only, uses the prediction of the direct CatBoost. The fourth, blend, mixes the residual and direct predictions with a weight selected from the discrete grid {0.00, 0.15, 0.30, 0.45, 0.60, 0.75, 1.00} by minimizing the RMSE on the internal validation set. The strategy chosen for testing the fold is the one with the lowest validation RMSE, subject to a minimum relative improvement of 0.5% over base only, which prevents marginal switches [28].

Three additional strategies were also used. First, recency weighting during training, with monotonically increasing linear weights (power exponent of 1.25), assigning more importance to recent observations. Second, an ensemble with two initial random configurations (using seeds 42 and 52) in the final stage, with arithmetic averaging of the outputs to reduce the variance inherent to stochastic boosting. Third, a seasonal baseline computed as the mean of lags 364 and 365 was provided as an additional attribute to the residual and direct models.

2.2.6. Walk-Forward Validation

The validation scheme adopted is strictly prospective, known as walk-forward with a rolling test window [22,29]. From January 1, 2018, consecutive bimonthly folds were defined, totaling 30 evaluation windows over the 2018 to 2022 period. For each fold, the training set covers the entire history before the test start date, and is divided into effective training (up to 75 days before the test start) and internal validation (the last 75 days, configured in the constant INNER_VAL_DAYS), used to select the predictive strategy and to calibrate the blend weight. After this selection, the models are refit with the full available training set and applied to the test fold.

Prediction over the test set is performed strictly recursively, in line with the realistic operational scenario of public health surveillance. For each new day in the window, three operations occur in sequence. First, the attribute matrix for the day is built from the observed history available up to the previous day. Second, the prediction is generated by the selected model. Third, the true value observed on the day just predicted is incorporated into the history, becoming available for the computation of lags and rolling windows in the following days. This one-step-ahead with feedback regime is more conservative and more realistic than schemes that rely solely on previous predictions to feed future lags [23].

2.2.7. Performance Metrics and Comparison Baselines

For each fold and for the annual aggregate, five complementary metrics were computed: mean absolute error (MAE), root mean squared error (RMSE), coefficient of determination (R^2), mean

 doi:10.20944/preprints202606.1217.v1

absolute percentage error (MAPE), and root mean squared logarithmic error (RMSLE). This set was adopted because each metric expresses a distinct aspect of performance. MAE provides a direct interpretation in the variable unit. RMSE penalizes large errors more strongly. R^2 measures the proportion of explained variance. MAPE expresses the error on a percentage scale. RMSLE attenuates the impact of extreme values through the logarithmic transformation [30].

Three comparison baselines were evaluated. The weekly baseline predicts, for each day, the value observed seven days earlier, capturing the strong weekly autocorrelation typical of hospital demand. The improved baseline corresponds to the ElasticNet component of the proposed model, equivalent to the first component of the hybrid architecture. The seasonal baseline relies on the mean of lags 364 and 365. The relative gain of the model (skill) was computed as (RMSE_baseline minus RMSE_model) divided by RMSE_baseline [31].

2.2.8. Interpretability Through SHAP Values

In order to discuss the substantive role of each exposure, an additional CatBoost model was trained, restricted to the twelve raw environmental and meteorological variables (without lags, anomalies, or interactions), with the residuals of the base model as the target. This restriction is justified for two reasons. First, the goal of the interpretation is to answer the central epidemiological question, that is, which environmental exposures most influence respiratory hospitalizations, and not the methodological question of which derived attributes the model uses most. Second, derived attributes are highly collinear with each other, which hinders direct epidemiological reading of the results [32].

Based on this model, SHAP values were computed per observation and summarized in four visualizations. The first is the mean absolute SHAP value, which ranks variables by the average magnitude of their contribution. The second is the CatBoost PredictionValuesChange importance plot, which measures the average impact of each variable on predictions. The third is the beeswarm plot, which shows the full distribution of SHAP values for each variable and reveals the direction and intensity of the effect. The fourth is the dependence plot of the two most important variables, which highlights the functional form of the relationship between the variable value and its effect on the prediction.

2.2.9. Computational Reproducibility

The entire implementation was carried out in Python 3.12, using the libraries pandas, numpy, scikit-learn, catboost, and matplotlib. Computational reproducibility was ensured by two complementary strategies. The first is the fixation of a global random seed (RANDOM_SEED = 42), propagated to all stochastic components of the pipeline, including the ElasticNet estimator and the CatBoost algorithm. The final CatBoost ensemble uses two distinct random initializations (seeds 42 and 52). The second is the centralization of hyperparameters in the CURATED_CFG dictionary of the source code, which enables future adjustments and replications in other Brazilian municipalities by editing a small number of code points [33].

Each run produces a standardized set of audit artifacts saved in a folder identified by a timestamp, including files with per fold metrics, concatenated daily predictions, annual metrics, the configuration actually used, and figures in PNG (300 dpi) and vector PDF format. This organization allows public auditing of the results, the reanalysis of any single fold, and the full reproduction of the study by third parties from the same input file.

2.3. Study 2: Digital Biomarkers via the Respire Bem System

2.3.1. The Respire Bem System

The Respire Bem system [17] was designed as a digital tool to support the complementary treatment of pediatric asthma through diaphragmatic breathing. The system implements a graphical interface suited to the pediatric age range and instruments the automatic collection of data during

 doi:10.20944/preprints202606.1217.v1

the exercise, recording the frequency of use, the duration and completeness of the breathing cycles, and the subjective state self-reported by the patient before and after each session. Although pharmacotherapy remains the standard treatment, complementary therapies such as diaphragmatic breathing have shown efficacy in reducing symptoms and anxiety levels, which reinforces the role of the system as a remote adherence instrument.

The scientific basis that supports the intervention comes from controlled clinical evidence, in particular the findings of Fernandes [16], who validated the efficacy of diaphragmatic breathing in a controlled environment. The clinical data indicated a significant improvement in the Asthma Control Test (ACT equal to 23.06 ± 1.62 in the intervention group) and a reduction in salivary cortisol levels (0.11 μg/dL in the intervention group) compared to the control group. The Respire Bem system therefore operates as an interface that transposes the clinical technique to the patient's daily routine.

2.3.2. ETL Pipeline and Inclusion Criteria

After daily use by patients over an average period of 12 months, a database was consolidated from the production environment through SQL queries on the MySQL database, totaling 1,698 exercise records from 102 distinct pediatric patients. To ensure the fidelity of the predictive models, an Extraction, Transformation, and Loading (ETL) pipeline was applied in three steps. The first step was the engineering of attributes derived from raw variables, including frequency of use, completeness of cycles, and temporal regularity of sessions. The second step was the normalization of scales. The third step was filtering by clinical and demographic criteria, which retained 913 qualified records from 84 patients, restricted to the pediatric age range from 6 to 16 years (mean age of 10.7 years; mean of approximately 11 qualified sessions per patient, range 1 to 47). The final feature set comprised the variables score_sentimento (self-reported sentiment score, ordinal scale), idade (age, in years), anos_asma (years with asthma), segundos_executados (effective seconds of exercise per session), as well as engineered attributes for adherence frequency, cycle completeness, and session regularity.

2.3.3. Evidence-Based Synthetic Data Simulation

Given the disparity between the volume of behavioral data collected by the system and the clinical and biological data that require laboratory tests, an Evidence-Based Synthetic Data Simulation strategy was adopted. For each of the 913 qualified records, a simulated ACT score and a simulated post-intervention salivary cortisol value were generated by sampling from Gaussian distributions calibrated on the trial of Fernandes [16] (intervention group: ACT mean = 23.06, standard deviation = 1.62; salivary cortisol mean = 0.11 μg/dL). The location parameter of each Gaussian draw was modulated by a monotonic function of the patient adherence profile, so that records with higher adherence and higher self-reported sentiment received simulated outcomes closer to the intervention-group mean, and records with lower adherence received simulated outcomes closer to the control-group mean reported in the same trial. This strategy preserves the statistical signal observed in the controlled clinical evidence and exposes the predictive pipeline to the variability expected in a naturalistic pediatric population. This approach is consistent with the methodological principle of complementarity of sources in digital health, in which controlled clinical evidence at small scale supports the interpretation of large scale behavioral data collected in naturalistic settings. The implications of this strategy, particularly for the interpretation of the performance coefficients, are discussed in Section 4.7 (Limitations).

2.3.4. Predictive Models: XGBoost and Random Forest

The computational modeling employed two algorithms well established in the literature. XGBoost is an optimized gradient boosting algorithm with L1 and L2 regularization and explicit handling of missing values [20]. Random Forest is an ensemble of decision trees built through bagging with random selection of attributes at each split. The choice was motivated by the strong

 doi:10.20944/preprints202606.1217.v1

ability of these methods to deal with heterogeneous tabular variables and by their interpretability through feature importance analysis. The validation followed a grouped train-test split at the patient level (70% training, 30% testing) that prevents leakage among observations of the same individual. Hyperparameters were tuned by grid search with 5-fold cross-validation inside the training partition. The confidence intervals of the performance metrics were estimated by non-parametric bootstrap with 1,000 resamples on the test partition. The focus of the evaluation is not limited to raw statistical accuracy, but also considers the ability of the algorithm to discriminate, with adequate sensitivity, the simulated transition of the patient to a state of clinical improvement, in line with the exploratory and methodological nature of Study 2.

2.4. Ethical Aspects

Study 1 used only secondary data of public domain, aggregated at daily frequency and at the municipal scale, with no information that allows direct or indirect identification of patients. The hospital admission data were obtained from SIH/SUS via PCDaS/Fiocruz, which provides only aggregated counts by municipality, age range, and ICD-10 chapter, without identifying variables. The environmental (CETESB) and meteorological (INMET) data refer to measurements from fixed monitoring stations, which are also public and aggregated. The research falls within the situations that are exempt from review by an Ethics Committee, according to Resolutions number 466/2012 and 510/2016 of the National Health Council, in particular Article 1, sole paragraph, item III of CNS Resolution 510/2016, which exempts from registration and assessment by the CEP/CONEP system the research that uses information of public domain.

Study 2 was conducted under a protocol approved by the Research Ethics Committee in the Health Area (CEP) of FURG (approval number 3,482,646), with a Free and Informed Consent Form signed by the legal guardians of all pediatric participants. The study was additionally registered in the Brazilian Registry of Clinical Trials (ReBEC) under number RBR-3x8py3n. All ethical principles applicable to health research with primary and secondary data were observed, including respect for the original purpose of data collection, the preservation of anonymity in all analysis steps, and a commitment to the responsible disclosure of the findings.

3. Results

3.1. Study 1: Environmental and Meteorological Modeling

3.1.1. Exploratory Characterization of the Database

The consolidated database totaled 2,191 consecutive days, in agreement with the six complete years of the study period, considering the leap year of 2020. After the deterministic linkage and consolidation pipeline described in Section 2.2.2, the thirteen variables of the final daily database showed full completeness (no missing values), with all formatting inconsistencies resolved. The causal imputation protocol described in Section 2.2.4 was therefore implemented as a safety mechanism of the pipeline, ready to handle any future missing observations during operational use, but was not extensively triggered in the present analysis. The annual mean of HOSPCIDX remained relatively stable between 2017 and 2019 (148.3, 148.6, and 143.4 admissions per day, respectively), showed a marked decrease in 2020 (117.3 admissions per day, a reduction of approximately 21% with respect to the previous three years), a partial recovery in 2021 (129.7), and a return close to pre-pandemic levels in 2022 (144.3). This pattern is consistent with the well-documented effects of the social distancing measures adopted during the COVID-19 pandemic on the circulation of other respiratory viruses and on the postponement of elective admissions [34,38].

3.1.2. Overall Performance and Comparison with Baselines

Table 2 summarizes the performance of the proposed model and of the comparison baselines for the consolidated annual results from 2018 to 2022. The hybrid model consistently outperformed the

 doi:10.20944/preprints202606.1217.v1

three baselines in every year evaluated, with the largest gains over the seasonal baseline (skill between 32.1% and 56.3%). The gain over the ElasticNet improved baseline was modest but consistently positive, which shows that the CatBoost component adds incremental value to the regularized baseline. The mean metrics for the full period were MAE of 18.22, RMSE of 23.99, and R^2 of 0.675. The superiority of the hybrid model over each baseline was confirmed in a sensitivity analysis based on per-fold RMSE differences and non-parametric bootstrap resampling with 1,000 replicates, which yielded 95% confidence intervals that excluded zero in every comparison.

Table 2. Annual performance of the hybrid model compared to baselines, 2018 to 2022.

Year	MAE	RMSE	R^2	MAPE (%)	RMSLE	Skill vs Seasonal
2018	15.11	19.75	0.794	11.1	0.143	32.1%
2019	13.93	17.92	0.808	10.7	0.134	35.1%
2020	21.63	28.34	0.600	21.1	0.246	44.2%
2021	22.98	29.55	0.640	19.7	0.235	39.9%
2022	17.45	24.40	0.533	12.3	0.170	56.3%
Mean	18.22	23.99	0.675	15.0	0.186	41.5%

Source: prepared by the authors.

Figure 1 shows the observed and predicted monthly time series throughout the full test period. The model recovers the seasonal peaks and troughs adequately, with mild underestimation of the extreme peaks during the winters of 2020 and 2021, periods that coincide with the first waves of the COVID-19 pandemic. Figure 2 displays the hexagonal parity plot between observed and predicted values, with a strong concentration of points around the identity line.

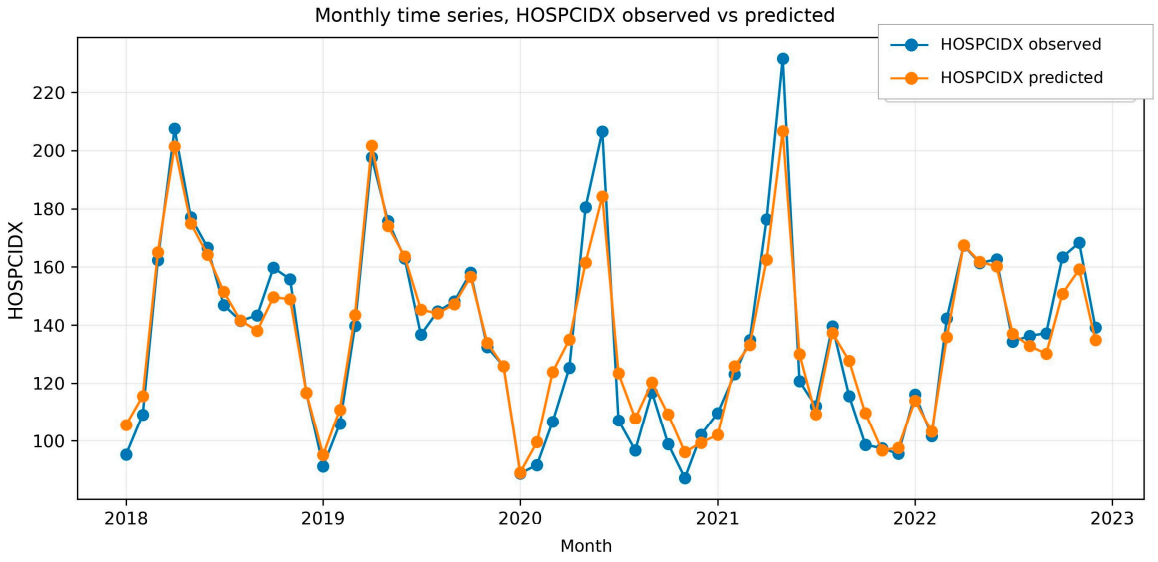


Figure 1. Monthly time series of observed versus predicted HOSPCIDX over the test period, 2018 to 2022. The model adequately captures the seasonal cycle and responds to the pandemic shocks of 2020 and 2021.

 doi:10.20944/preprints202606.1217.v1

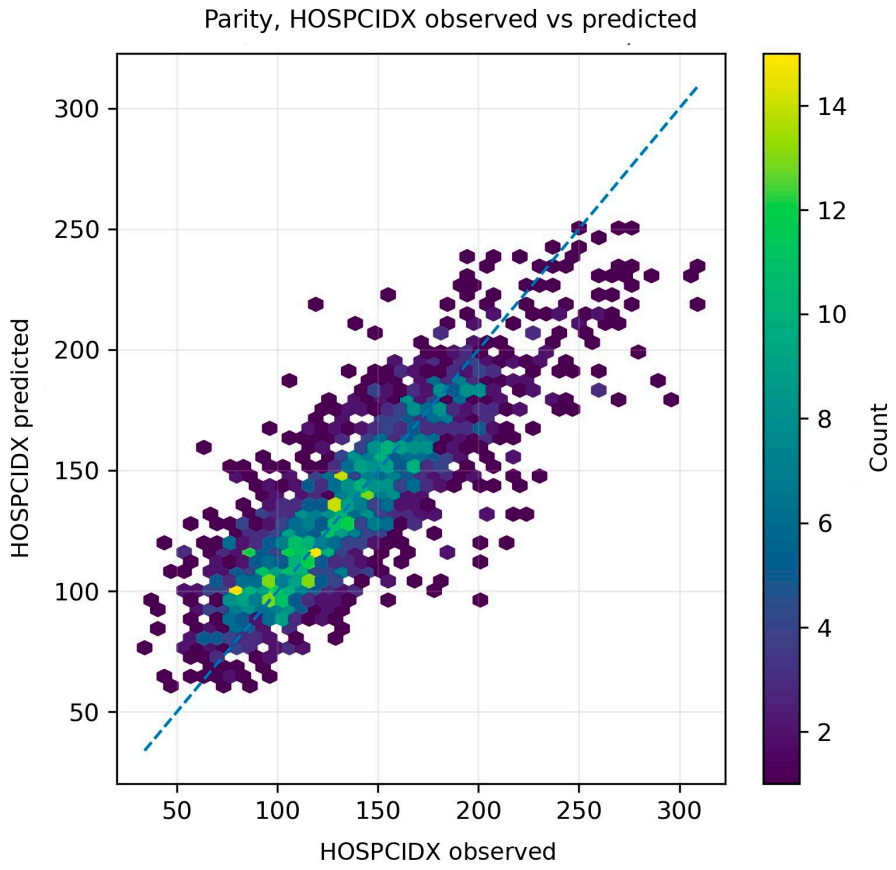


Figure 2. Hexagonal parity plot between observed and predicted HOSPCIDX. The concentration along the identity line shows no systematic bias across a wide range of values.

3.1.3. Per Fold Performance and Adaptive Strategy

Figure 3 presents the RMSE in each of the 30 bimonthly folds, comparing the proposed model with the three baselines. The proposed model remained equal to or better than the best baseline in nearly all folds, with localized degradation in folds 15, 16, and 22, which coincide with the most acute pandemic waves and with abrupt changes in the behavior of the hospital series. Figure 4 summarizes the predictive strategy selected for each fold based on internal validation. The blend strategy was selected in 23 of the 30 folds, which shows that the adaptive combination of the residual and direct components consistently outperforms the isolated use of any one of them. The direct only strategy was chosen in 4 folds, residual only in 2 folds, and base only in 1 fold.

 doi:10.20944/preprints202606.1217.v1

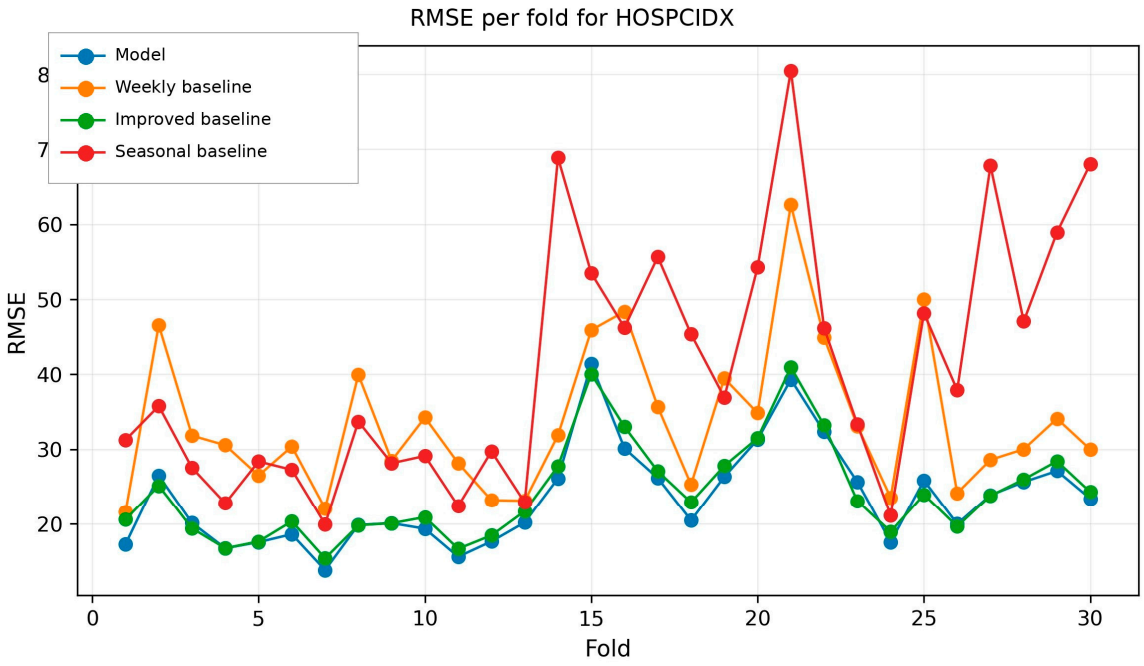


Figure 3. RMSE per fold for the proposed model and the three baselines (weekly, improved, and seasonal). The model maintains a consistent advantage, with localized degradation in the folds affected by the pandemic waves.

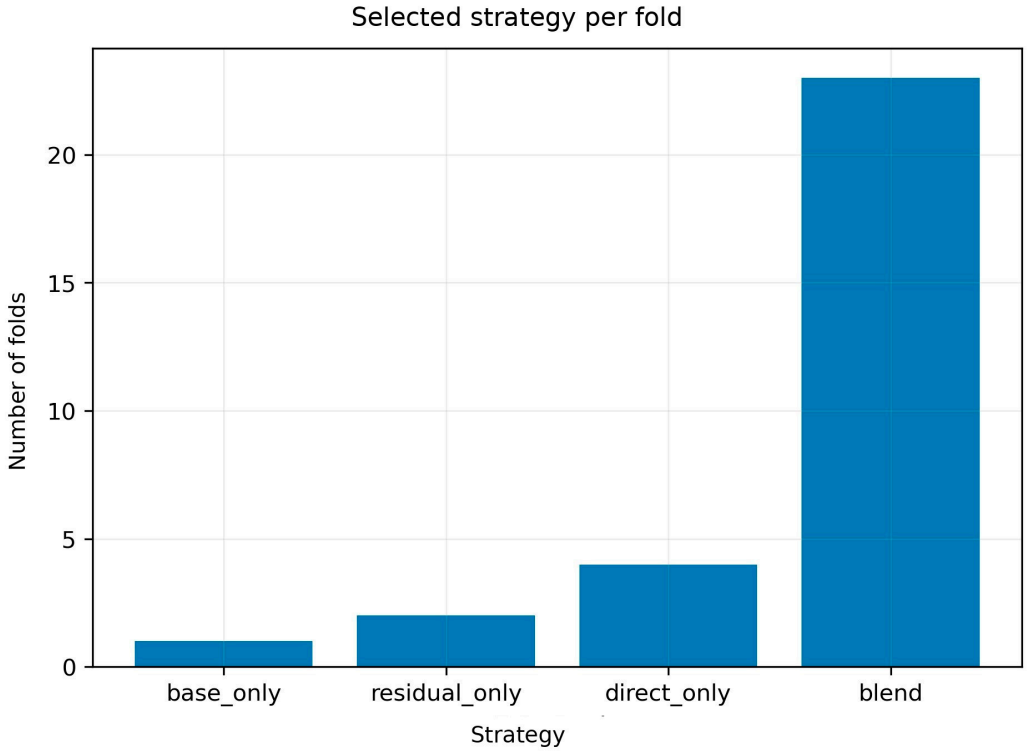


Figure 4. Distribution of the predictive strategy selected per fold. The adaptive blend was chosen in 23 of the 30 folds, which validates the hybrid approach.

Figure 5 presents the heat map of relative performance across the 30 walk-forward folds. The model maintained stable, high performance in most temporal windows (lighter cells), with localized degradation in folds 15, 16, and 22 (darker cells across RMSE, MAE, MAPE, and RMSLE). These folds coincide with the most acute pandemic waves and abrupt shifts in hospital demand, confirming that

 doi:10.20944/preprints202606.1217.v1

the loss of accuracy is concentrated in periods of structural change rather than distributed across the series.

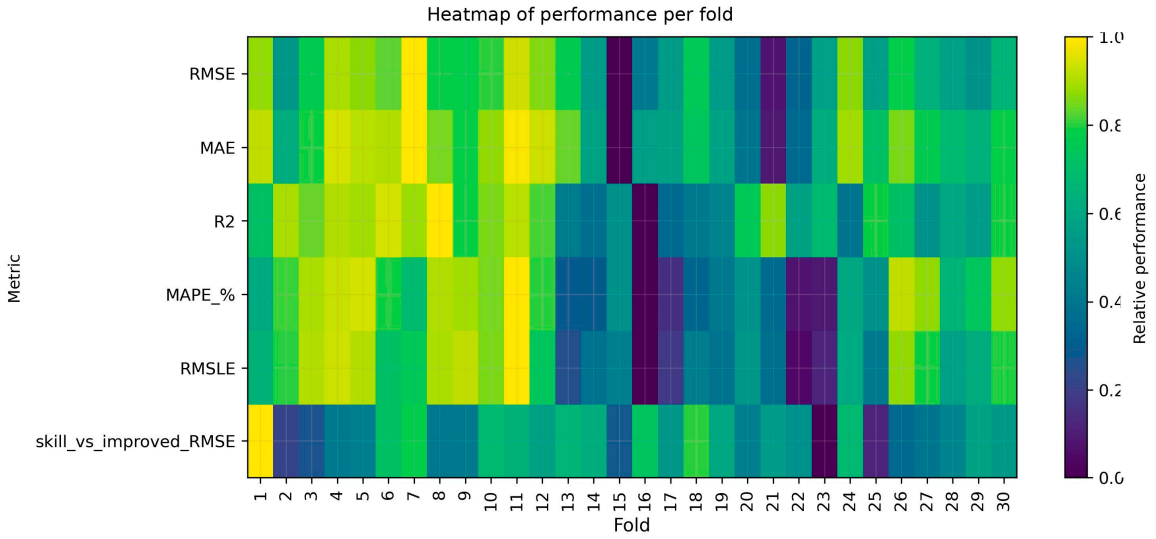


Figure 5. Heat map of relative performance across folds and metrics. Lighter cells indicate better performance; darker regions in folds 15, 16, and 22 coincide with COVID-19 pandemic shocks.

3.1.4. Annual Distribution of Errors

Figure 6 presents the distribution of the absolute error by year. The dispersion increased in 2020 and 2021, which is consistent with the rise in extreme events during the pandemic period, and returned partially to normal in 2022.

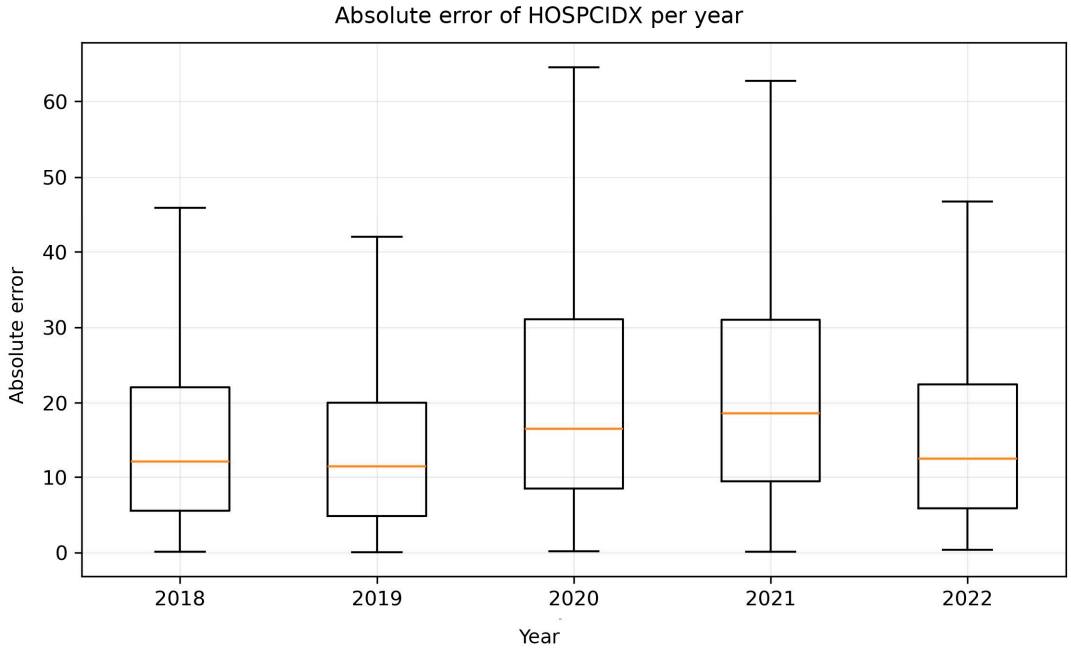


Figure 6. Annual distribution of the absolute error. The dispersion increases in 2020 and 2021 (pandemic period) and partially returns in 2022.

3.1.5. Interpretability: Importance and Direction of Environmental Drivers

 doi:10.20944/preprints202606.1217.v1

As detailed in Section 2.2.8, an auxiliary CatBoost model was trained, restricted to the twelve raw environmental and meteorological variables, exclusively for interpretability purposes. Figure 7 presents the PredictionValuesChange importance, while Figure 8 summarizes the mean absolute SHAP values. The two indicators partially converged. TEMPMED, PM2.5, and CO appeared among the most relevant variables in both rankings, whereas NO2 was more prominent in the SHAP ranking than in the PredictionValuesChange ranking. Figure 9, in turn, allows the observation of the direction and dispersion of the estimated effects.

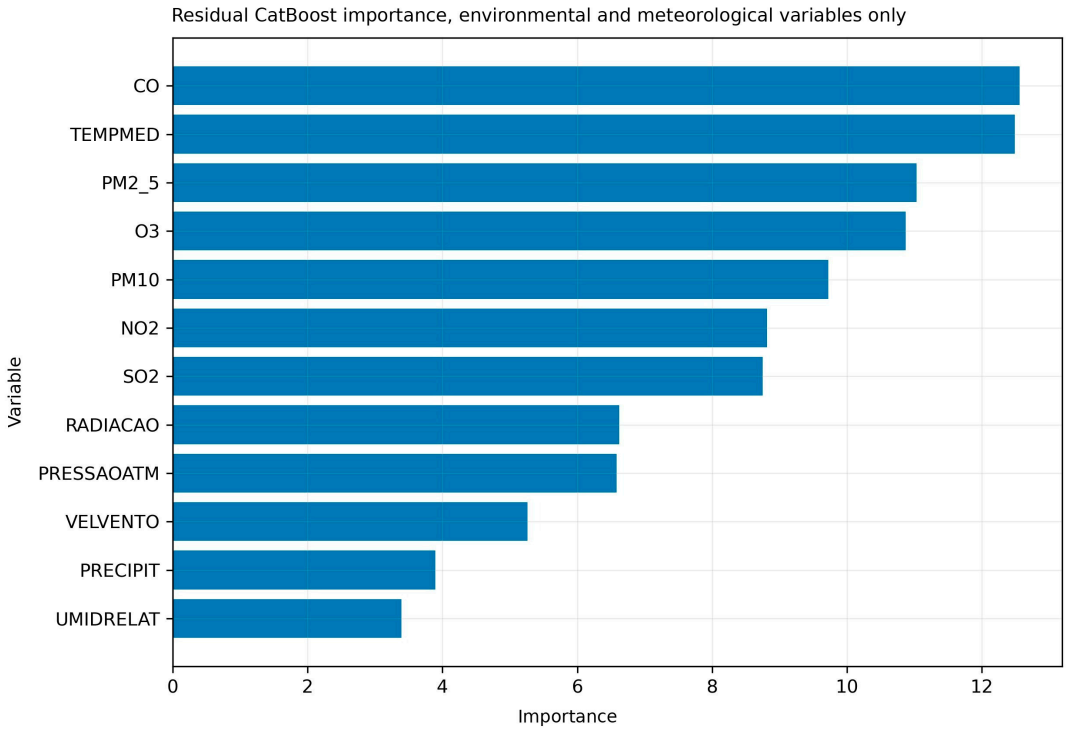
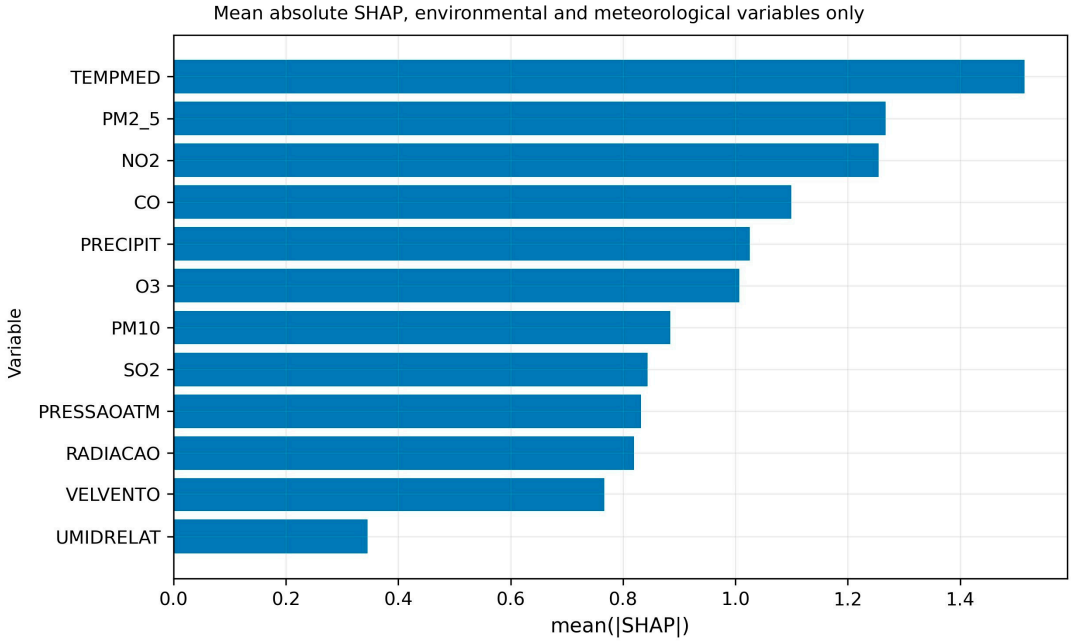


Figure 7. CatBoost importance (PredictionValuesChange) for raw environmental and meteorological variables in the residuals model.



 doi:10.20944/preprints202606.1217.v1

Figure 8. Mean absolute SHAP values for environmental and meteorological variables. TEMPMED, PM2.5, and NO2 dominate, followed by CO and PRECIPIT.

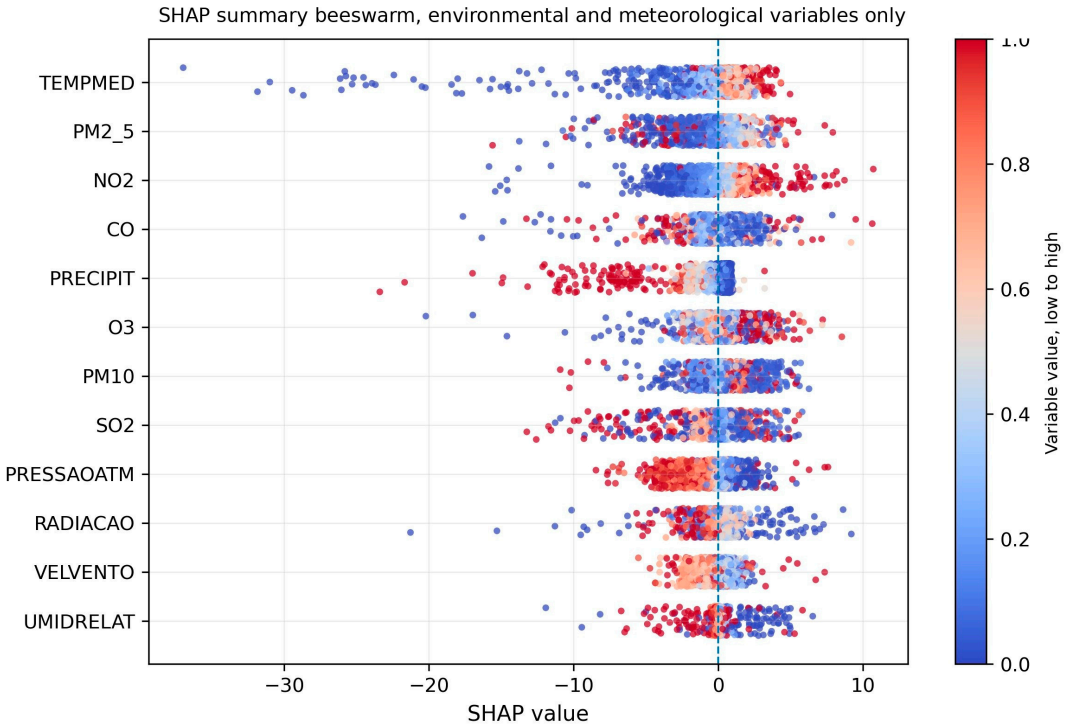


Figure 9. Beeswarm plot of SHAP values for the residuals model. Each point represents an observation; color encodes the variable value (blue, low; red, high). For TEMPMED, low values yield strongly negative SHAP contributions, reflecting a downward correction of the seasonal/linear baseline (the residual is most negative on cold days that the base model overestimates). Pollutant peaks (PM2.5, NO2, CO) yield positive SHAP contributions, indicating residual increases above the baseline.

Figures 10 and 11 present, respectively, the SHAP dependence plots for the two most influential variables. It is important to recall that the target of the interpretability model is the residual of the ElasticNet base component, which already encodes the linear and seasonal effect of temperature. The SHAP values therefore represent the additional adjustment of the CatBoost model on top of that baseline, rather than the raw association between temperature and hospitalizations. For TEMPMED (Figure 10), a strongly nonlinear relationship is observed, with negative SHAP contributions at temperatures below 14 °C, an abrupt drop below this threshold, and stabilization close to zero above 18 °C. The negative sign indicates that, on cold days, the base model tends to overestimate admissions and the residual model applies a downward correction. The non-linear shape is consistent with the well-documented effect of cold on the airways, particularly in pediatric and elderly populations, as reported by Hyrkas et al. [35], Achebak et al. [11], and Li et al. [36]. For PM2.5 (Figure 11), a biphasic pattern is observed, with modest positive contributions at typical concentrations (10 to 25 μg/m³) and negative contributions at extreme concentrations (above 30 μg/m³), a pattern that deserves discussion in Section 4.1.

 doi:10.20944/preprints202606.1217.v1

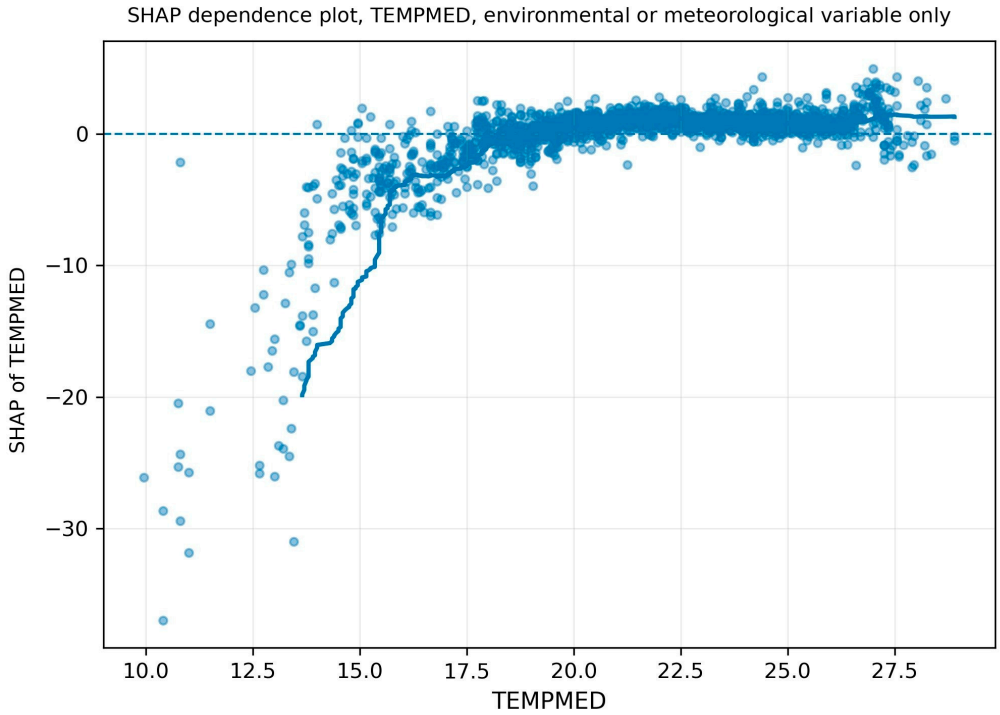


Figure 10. SHAP dependence plot for TEMPMED in the residuals model. The strongly negative SHAP values below approximately 14 °C indicate that, at low temperatures, the residual is below the prediction of the base seasonal/linear model. The base model already incorporates part of the seasonal cold-related rise in admissions, so SHAP captures the additional adjustment needed when the magnitude of that rise was overestimated by the baseline.

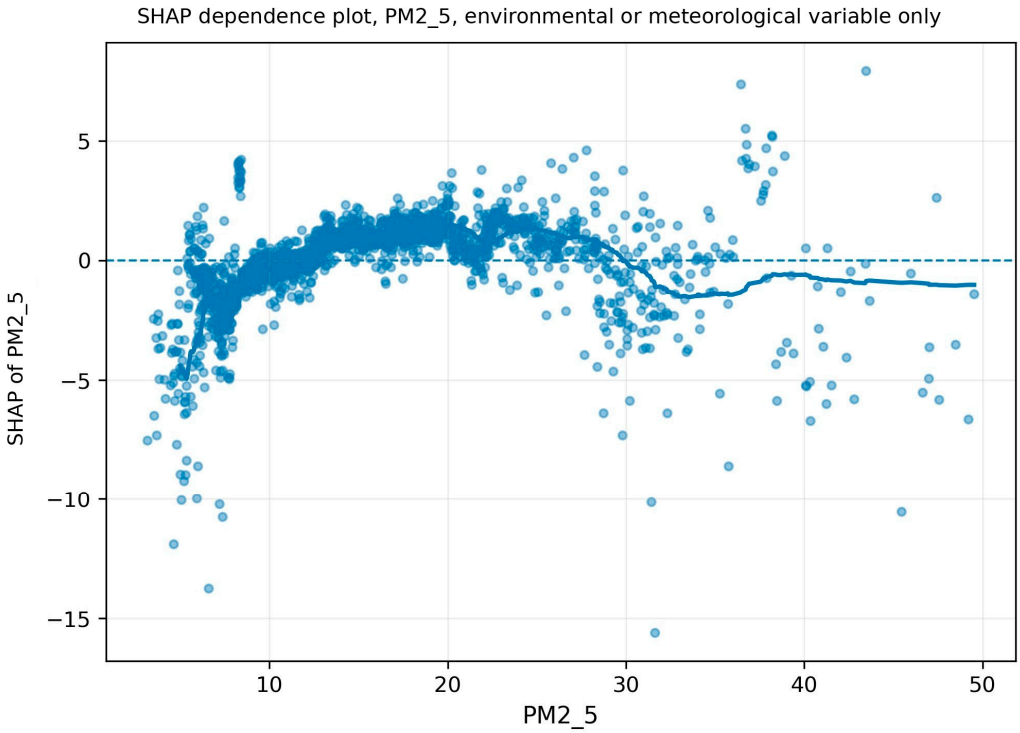


Figure 11. SHAP dependence plot for PM2.5. Biphasic pattern with positive contributions at typical levels and negative contributions at extreme concentrations.

3.1.6. Residual Diagnostics and Sensitivity Analysis

 doi:10.20944/preprints202606.1217.v1

The analysis of the autocorrelation of residuals (Figure 12) shows low to moderate values, on the order of 0.14 at lag 1 and 0.16 at lag 2, decaying progressively to about 0.03 at lag 30. This pattern indicates that the temporal structure was mostly captured by the lag components of the target series, but a short-term residual signal remains and could be improved by models with distributed lag effects, as discussed in Section 4.1.

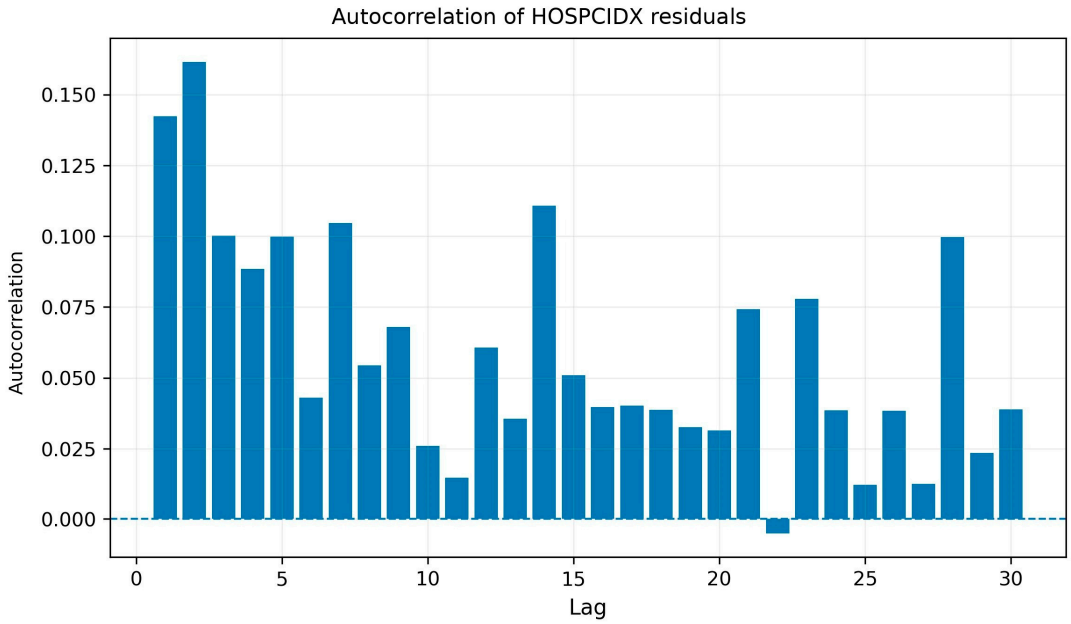


Figure 12. Residual autocorrelation of the model up to lag 30. The largest values (lags 1 and 2) suggest short-term residual persistence, with no pronounced seasonal pattern.

The robustness of the findings was additionally evaluated through a sensitivity analysis along four dimensions. The first was the variation of the temporal validation fold size by changing the FOLD_MONTHS parameter to 1, 2, and 3 months. The second was the variation of the internal validation window size by changing INNER_VAL_DAYS to 60, 75, and 90 days. The third was the inclusion or exclusion of the binary pandemic indicator. The fourth was the systematic exclusion of one environmental variable at a time (leave-one-feature-out), to identify which exposures contribute essentially to the model performance. In all tested variations, the ranking of variable importance remained stable (TEMPMED, PM2.5, NO2, and CO were always among the first five positions), and the gain of the model over the baselines remained positive, with variations of less than 5% in the mean test RMSE.

3.2. Study 2: Digital Biomarkers

3.2.1. Sample Characteristics

The qualified sample of the Respire Bem system totaled 913 records of breathing exercises from pediatric patients aged 6 to 16 years, with a mean age of 10.7 years. The 913 records resulted from the ETL pipeline applied to an initial raw database of 1,698 records, configuring a qualification rate of approximately 53.8%. This rate reflects the pattern observed in longitudinal studies of digital health with pediatric populations, in which a substantial share of records is discarded due to incomplete breathing cycles, interrupted sessions, age outside the inclusion criterion, or absence of self-reported subjective variables. The mean follow-up period was 12 months, which characterizes the longitudinal nature of the study.

The operationalization of the study, anchored in the controlled clinical evidence reported by Fernandes [16] and supported by the Respire Bem system [17] as an interface between clinic and daily routine, defines a scenario in which behavioral data collected in real use are paired with simulated

 doi:10.20944/preprints202606.1217.v1

clinical biomarkers calibrated to that prior evidence. This pairing supports the construct validity of the proposed design at an exploratory and methodological level, and is not equivalent to direct clinical validation.

3.2.2. Performance of the XGBoost and Random Forest Models

The XGBoost and Random Forest algorithms were evaluated for predicting two simulated clinical outcomes, generated by evidence-based synthetic simulation calibrated on the distributions reported in the clinical trial of Fernandes [16]. The two outcomes were the simulated Asthma Control Test (ACT) score and the simulated post-intervention salivary cortisol concentration. The choice of the two algorithms was motivated by their ability to operate with heterogeneous tabular data and by their native interpretability through feature importance, which is particularly relevant in research that dialogues with the clinical and biomedical literature.

For the ACT outcome, both algorithms achieved R^2 above 0.79, with XGBoost slightly ahead (R^2 = 0.80; RMSE = 0.85; MAE = 0.64; see Table 3). The models were able to discriminate different levels of asthma control from the digital biomarkers of adherence and self-reported sentiment, supporting their use as computational proxies of the expected clinical response. For the simulated salivary cortisol outcome, the models reproduced the expected trend of biomarker reduction associated with consistent adherence to the diaphragmatic breathing protocol, in line with the clinical evidence of stress reduction associated with the intervention.

Table 3. Predictive performance of the models for the simulated clinical outcomes (ACT and salivary cortisol, generated by evidence-based synthetic simulation). 95% confidence intervals obtained by non-parametric bootstrap with 1,000 resamples.

Outcome	Model	R^2	RMSE	MAE
ACT	XGBoost	0.80 [0.77, 0.83]	0.85 [0.80, 0.90]	0.64 [0.60, 0.68]
ACT	Random Forest	0.79 [0.76, 0.82]	0.84 [0.79, 0.89]	0.63 [0.59, 0.67]
Cortisol	XGBoost	0.78 [0.75, 0.81]	0.03 [0.02, 0.04]	0.02 [0.01, 0.03]
Cortisol	Random Forest	0.78 [0.75, 0.81]	0.03 [0.02, 0.04]	0.02 [0.01, 0.03]

Source: prepared by the authors.

Table 3 shows comparable performance between the two models, with coefficients of determination (R^2) above 0.75 in all scenarios and narrow confidence intervals. XGBoost had a slightly higher R^2 for the simulated ACT outcome (R^2 = 0.80 versus 0.79), whereas Random Forest showed marginally lower RMSE and MAE (0.84 and 0.63 versus 0.85 and 0.64). For the simulated salivary cortisol, both algorithms produced essentially identical metrics. The absolute error magnitudes are small relative to the real scales of the outcomes (the ACT theoretical scale spans 5 to 25, with values observed in the sample concentrated in the range of approximately 10 to 25; cortisol is typically between 0.05 and 0.40 μg/dL). The implications of these coefficients, given the strategy of synthetic simulation of the clinical data, are contextualized in Section 4.7 (Limitations).

3.2.3. Importance of Behavioral Attributes

To ensure transparency of the model decisions and to validate the clinical relevance of the collected data, the technique of Explainable Artificial Intelligence (XAI) was used. The importance of the attributes was quantified by the Information Gain (Gain) metric of the XGBoost algorithm, which measures the relative contribution of each variable to the reduction of the global error in the decision trees. This approach allows the identification of which digital biomarkers have the greatest discriminative power in predicting asthma control. Figure 13 presents the resulting hierarchy of importance.

 doi:10.20944/preprints202606.1217.v1

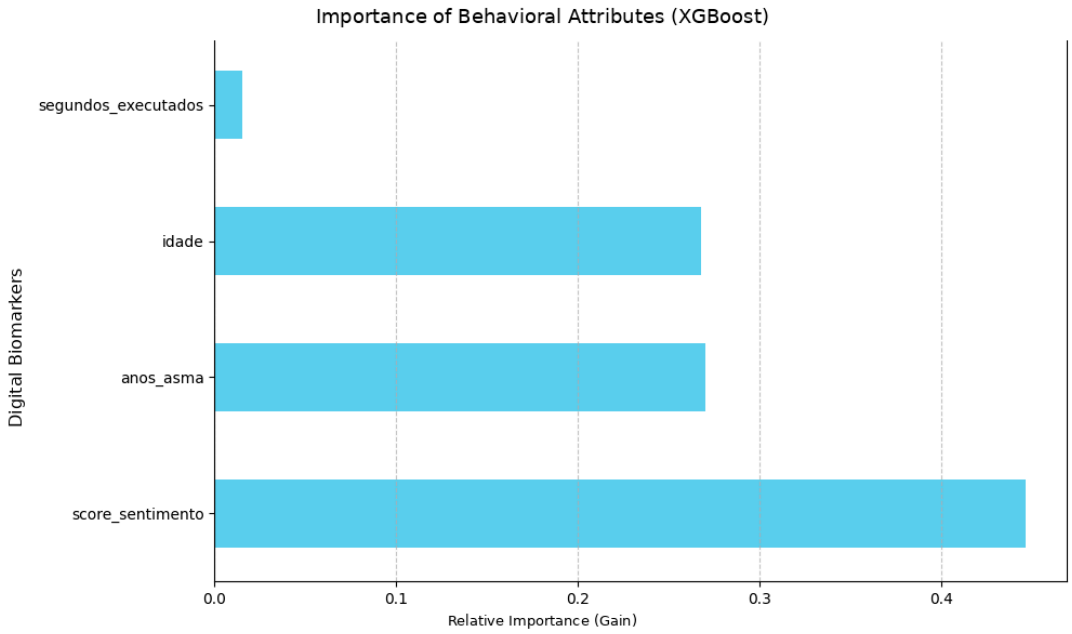


Figure 13. Hierarchy of importance of digital biomarkers and clinical variables in the predictive modeling of Study 2 (XGBoost, Gain metric). Score_sentimento: the score in self-analisys; Anos_asma: years with asthma; Idade: age; Segundos_executados: time that the patient uses the Respire Bem system.

The analysis of Figure 13 shows that the attribute score_sentimento (sentiment score), which corresponds to the subjective perception of well-being self-reported by the patient, emerged as the most impactful predictor in the model, with a relative importance of approximately 0.45. Next, with similar importance (around 0.27 each), appear the variables anos_asma (years with asthma, capturing disease chronicity) and idade (age). The variable segundos_executados (seconds performed), which measures the effective execution time of the breathing exercise, showed the lowest relative importance (close to 0.02), an inversion of hierarchy compared to what could be expected from a strictly quantitative view of the exercise.

This pattern has two main, non-exclusive interpretations. First, the immediate emotional response of the patient to diaphragmatic breathing acts as a sensitive signal of autonomic modulation, capturing therapeutic effects that strictly quantitative metrics of duration and completeness may omit. Second, there are signs of the existence of a behavioral efficacy threshold: once the minimum time required for the completeness of the breathing cycle is reached, the marginal variation of additional seconds becomes less decisive for the clinical outcome than the subjective state and the clinical profile of the patient (age and duration of chronicity). This finding is coherent with the emerging literature in digital health, where patient-reported outcomes are combined with usage logs to form a holistic view of treatment [18]. The consistency between digital behavior and clinical profile reinforces that the Respire Bem system operates not only as a data repository, but as a clinical audit instrument capable of prioritizing subjective signals of improvement in the prediction of asthma control.

Figure 14 presents the relationship between the projected (simulated) ACT values (x-axis) and the values predicted by the XGBoost model (y-axis). The dispersion shows reasonable adherence to the identity line (y = x), with points distributed across the full ACT range (approximately 10 to 25), which supports the interpretation of the coefficients reported in Table 3 ($R^2 = 0.80$; MAE = 0.64). Unlike patterns in which predictions converge to the global mean of the response variable, the distribution observed in Figure 14 shows that the model discriminates different levels of asthma control from the digital biomarkers of adherence and sentiment, which is consistent with the hypothesis that patterns of digital interaction serve as useful proxies of the expected clinical response. The cautious interpretation of these findings, especially regarding the generalization to real clinical populations, is discussed in Section 4.7 (Limitations).

 doi:10.20944/preprints202606.1217.v1

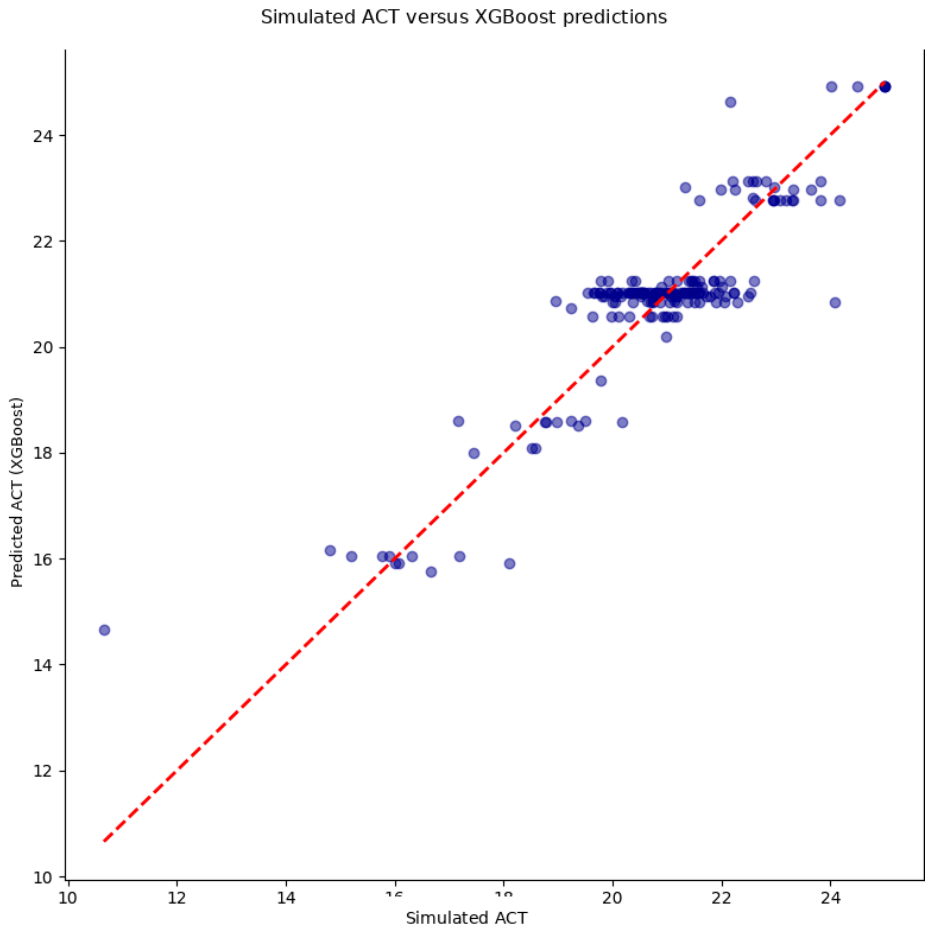


Figure 14. Scatter plot between the simulated ACT values (synthesized from the distributions reported by Fernandes [16]) and the values predicted by XGBoost. The red dashed line represents the identity y = x.

4. Discussion

4.1. Environmental Drivers and Biological Plausibility

The findings of Study 1 are consistent with the recent epidemiological literature on air pollution, meteorology, and respiratory disease. It is worth restating that the SHAP analysis was applied to the residuals of the ElasticNet base model rather than to raw counts. Consequently, the contributions reported below describe the additional non-linear corrections captured by the CatBoost component, not the raw epidemiological association between exposures and hospitalizations. The dominance of TEMPMED as the most influential residual driver reflects the well-known role of temperature in modulating respiratory risk, especially in pediatric and elderly populations, where bronchial hyperreactivity can be triggered by exposure to cold and dry air [35]. The nonlinear relationship observed in Figure 10, with a sharp change in the SHAP contribution around 14 °C, suggests the existence of a possible operational thermal threshold, with implications for public health alert systems [11,36].

The high importance of PM2.5, NO2, and CO converges with the current consensus that fine particulate matter and the markers of vehicular combustion dominate the anthropogenic component of urban respiratory risk [5,12]. In densely populated centers, these three pollutants together reflect the intensity of traffic, the burning of fossil fuels, and industrial processes, articulating the physical environment with the social component of respiratory risk [8,37].

The biphasic pattern observed for PM2.5 (Figure 11), with a positive contribution at typical levels and a negative contribution at extreme concentrations, deserves careful interpretation. This behavior may reflect three overlapping phenomena. First, behavioral warning effects, where individuals with

 doi:10.20944/preprints202606.1217.v1

asthma reduce their outdoor exposure on peak pollution days. Second, correlation with specific atmospheric episodes, such as thermal inversions associated with dry periods and lower population circulation. Third, delays in seeking care during extreme events, which manifest as hospitalizations in subsequent days rather than on the peak day itself. The dedicated investigation of this pattern, with distributed lag windows, is a promising direction for future work [13,36].

4.2. Pandemic Shock and Model Robustness

The drop in R^2 during 2020 and 2021, from approximately 0.80 to 0.60, is compatible with the impact of the COVID-19 pandemic on the dynamics of respiratory hospitalizations. During this period, three relevant phenomena occurred simultaneously. First, an abrupt increase in admissions for respiratory causes attributable to COVID-19 itself. Second, changes in the usual patterns of care seeking. Third, non-pharmacological measures such as lockdowns and the use of masks altered the circulation of other respiratory pathogens [34,38]. The pandemic flag incorporated as a feature absorbed part of this heterogeneity, but did not eliminate it completely, which represents an expected limitation in time series affected by abrupt structural changes.

Despite this degradation, the model maintained an advantage over all baselines in every fold and year evaluated, demonstrating robustness to regime changes, a desirable property in public health tools that will inevitably be exposed to future exogenous shocks.

4.3. Digital Biomarkers as Remote Clinical Observers

The findings of Study 2 support the technical feasibility of operationalizing digital biomarkers based on behavioral and subjective variables captured by the Respire Bem system. The ability of the XGBoost and Random Forest models to learn patterns among adherence metrics, clinical profile, self-reported sentiment, and clinical outcomes simulated from prior evidence positions the system as a candidate tool for exploratory remote clinical monitoring, pending prospective clinical validation. The originality of Study 2 does not lie in the revalidation of diaphragmatic breathing as an intervention, already documented previously [16], but in showing that patterns of digital interaction, in particular the self-reported subjective state, can serve as computational proxies of the expected clinical response. In other words, the computational model operates as a support instrument for remote follow-up of the complementary intervention for asthma, an approach aligned with contemporary trends in pediatric digital health.

The most clinically relevant and counterintuitive finding emerges from the analysis of attribute importance (Figure 13): the subjective attribute (score_sentimento, sentiment score) outperforms, by a wide margin, all strictly objective adherence metrics (in particular, segundos_executados, seconds performed, which appears in the last position). This result has three main implications. First, it suggests that, once a minimum threshold of technical exercise execution is reached, additional gains in duration do not translate proportionally into improvements in clinical outcome, in line with the notion of saturated dose response known in behavioral interventions. Second, it values the pediatric self-report as first line predictive information, contrary to the tendency to underestimate this source for being considered more susceptible to bias. Third, it repositions the role of structural clinical variables (age and duration of disease) as relevant modulators of the response, a hierarchy distinct from that observed in models based solely on aggregated behavioral logs. Together, these results suggest that adherence support applications should prioritize the systematic capture of the subjective state of the patient, ideally combined with objective usage metrics and the clinical profile, building a multidimensional set of digital biomarkers. The ability to detect an early decline in self-reported sentiment can trigger targeted interventions in high risk patients, with potential to reduce exacerbations and hospital costs.

4.4. Integrated Framework: Macro-Micro Articulation

 doi:10.20944/preprints202606.1217.v1

The combination of the two studies suggests a two-layer analytical architecture. At the population layer, the environmental model anticipates periods of elevated risk of hospitalizations and provides input for targeted alerts. At the individual layer, the digital biomarkers model identifies specific patients with signals of declining adherence or worsening subjective state, enabling proactive interventions. When integrated, these two layers form a surveillance and clinical support system that articulates population-level prevention with individualized care.

Operationally, the articulation can occur in three steps. First, the environmental model signals, with a variable lead time, periods of elevated risk for the municipality of São Paulo or for the monitored area. Second, the Respire Bem system identifies, among its users, patients with recent signs of decline in the self-reported subjective state and in adherence. Third, targeted interventions, such as reinforcement messages, proactive contact from the clinical team, or therapeutic adjustments, can be offered to the patients with the greatest overlap between environmental and individual risk. This approach is in line with contemporary trends of integration between public health and personalized medicine [24,26].

4.5. Comparison with Other Studies in the Field (Study 1)

The results of Study 1 are compatible with the recent literature on the prediction of respiratory admissions through machine learning. Cabral-Miranda et al. [24], in a Brazilian pediatric population, reported R^2 values between 0.55 and 0.72 using an XGBoost based architecture with climatic and pollution factors. Barnett-Itzhaki et al. [25], in an Israeli pediatric population, achieved similar performance in predicting hospitalizations associated with air pollution and humidity. Bochenek et al. [14], in a Polish study with a nine-year time series, identified PM10 and NO2 as the main predictive factors of emergency COPD admissions, a finding consistent with the present work.

In direct comparison with these studies, the present work offers four methodological differentials. First, walk-forward validation with 30 bimonthly folds, which is more conservative than the static train-test splits predominant in the literature. Second, a hybrid architecture with adaptive strategy selection, instead of a monolithic one. Third, SHAP interpretability restricted to raw variables, which avoids collinearity among derived attributes, in line with the approach proposed by Lundberg et al. [32]. Fourth, simultaneous coverage of the pre-pandemic, pandemic, and post-pandemic periods, which allows robust analysis of regime change, a dimension little explored in the literature.

Table 4. Comparison between the present work and related machine learning studies on respiratory hospitalizations.

Study	Country	Period	Model	R^2	Main predictors / gap
Cabral-Miranda et al. [24]	Brazil	2014 to 2018	XGBoost + clinical features	0.55 to 0.72	Pediatric care; limited environmental coverage
Barnett-Itzhaki et al. [25]	Israel	2010 to 2017	Random Forest	0.65 to 0.75	Air pollution and humidity; single setting
Bochenek et al. [14]	Poland	2012 to 2019	Generalized linear models	n.r.	PM10, NO2 dominant; COPD only
Present work	Brazil	2017 to 2022	Hybrid ElasticNet + CatBoost (residual + direct + blend)	0.675 (mean)	Walk-forward over 30 folds; SHAP on raw variables; pandemic regime

Source: prepared by the authors. n.r., not reported.

4.6. Comparison with Other Studies in the Field (Study 2)

 doi:10.20944/preprints202606.1217.v1

When comparing the results of this study with the survey carried out in the systematic review by Ferreira et al. [18], an expansion of the use of machine learning applied to asthma is observed. While most of the reviewed works focus on supporting clinical diagnosis and predicting crises based on hospital data or environmental sensors, Study 2 directs modeling toward adherence and the expected clinical response to a specific non-pharmacological intervention, diaphragmatic breathing. This approach helps fill a gap identified in that review, which pointed to the scarcity of predictive models focused on the monitoring of adherence and the potential efficacy of complementary therapies through digital interventions.

Unlike the traditional methodologies cited in the literature, which use static data from medical records, this study employs digital biomarkers extracted from real interaction logs of the Respire Bem system. While the Machine Learning models discussed in Ferreira et al. [18] show high accuracy in diagnosis through spirometry, the exploratory predictive efficiency observed here lies in the ability of the algorithms (XGBoost and Random Forest) to reproduce simulated clinical (ACT) and biological (cortisol) outcomes calibrated from prior evidence. This transition from the use of Machine Learning as a diagnostic tool to its use as a remote clinical audit instrument represents an advance in the personalization of pediatric asthma treatment.

4.7. Limitations

Although the two studies form an integrated framework for asthma management, it is important to acknowledge that they present different natures of evidence. Study 1 is based on real, secondary, population level data from official sources of health, air quality, and meteorology, with prospective temporal validation through walk-forward. Study 2, in turn, demonstrates the computational feasibility of digital biomarkers based on real behavioral data from the Respire Bem system, but uses clinical outcomes simulated from prior evidence. Therefore, its results should be interpreted as an exploratory and methodological stage that still depends on prospective validation with direct clinical and biological measurements.

The following limitations of Study 1 are acknowledged. First, the design is ecological, which does not allow individual-level causal inference. Second, exposures were aggregated at the municipal scale, ignoring intra-urban heterogeneity and indoor exposure [27]. Third, the residual autocorrelation at lags 1 and 2 (about 0.14 to 0.16) suggests room for incremental improvements in the temporal component, possibly through distributed lag non-linear models [13]. Fourth, the response variable aggregates all admissions for chapter X of the ICD-10, without separating asthma (J45, J46) from chronic obstructive pulmonary disease (J44) or other respiratory conditions, which limits clinical specificity. Fifth, the analysis does not include data on influenza vaccination coverage, school calendars, or aeroallergen levels (pollen and fungal spores), which are known modulators of respiratory hospitalizations and could refine predictive accuracy in future versions.

Sixth, and of central relevance for the interpretation of Study 2, the clinical outcomes (ACT and salivary cortisol) were generated by evidence-based synthetic simulation, using the normal distributions and standard deviations obtained in the controlled clinical trial of Fernandes [16]. This strategy, although methodologically justified by the unavailability of laboratory tests for the entire naturalistic population, tends to produce outcomes with less heterogeneity than that observed in real clinical populations. As a consequence, the coefficients of determination reported in Table 3 (R^2 of 0.80 for ACT and 0.78 for cortisol) should be interpreted as indicators of the internal consistency of the predictive pipeline and of the ability of the model to reproduce the signal contained in the reference distribution, and not as direct estimates of performance in prospective clinical use. The dispersion observed in Figure 14, with reasonable adherence to the identity line across the full ACT range (10 to 25), shows that the model does not collapse to the mean of the response variable, but the confirmation of the clinical generalization of these findings depends on validation in an independent population with direct laboratory measurements.

Seventh, derived from the previous limitation, prospective validation of the model in an independent clinical population is required, with direct application of the Asthma Control Test and

 doi:10.20944/preprints202606.1217.v1

laboratory measurement of salivary cortisol, before any use as therapeutic decision support. Eighth, the generalization of the findings of Study 1 to other Brazilian metropolitan regions needs confirmation in multicenter studies, given the specificity of the urban environmental ecosystem of São Paulo.

4.8. Practical Implications and Future Work

The practical implications of the findings extend in three directions. In public health, the pipeline can support short-term daily surveillance and alert systems linked to the municipal hospital contingency plan. The predictive horizon validated in this work is one-step-ahead with recursive updating from the previous observed day, which corresponds to a daily surveillance regime; multi-week operational forecasting would require a dedicated multi-step-ahead validation that is not part of the present scope. In personalized medicine, the integration with digital devices, such as Respire Bem, enables interventions targeted at higher risk patients during periods of critical environmental exposure. In environmental policy, the SHAP quantification by pollutant provides empirical evidence for discussions about exposure limits and low emission zones.

Promising future directions for Study 1 include (i) decomposition of the target variable by clinical subgroups (asthma, ICD J45 and J46, versus COPD, ICD J44), enabling specialized models; (ii) the incorporation of models with distributed lag effects (DLNM) to strengthen causal inference and attenuate residual autocorrelation; (iii) the prospective coupling of the two studies in an outpatient clinical population, articulating environmental alerts with adherence reinforcement interventions; (iv) the extension of the hybrid architecture to other large Brazilian metropolitan regions, in collaboration with Municipal Health Departments; (v) the assessment of climate change scenarios (SSP-RCP) on the projection of respiratory hospital demand up to 2050; and (vi) the calibration of the model into operational decision thresholds, such as the probability of exceeding the 90th percentile of admissions over the next 7 to 14 days, enabling direct use in hospital surge planning.

Promising future directions for Study 2 include (i) the improvement of the algorithmic intelligence through sequential models, such as recurrent neural networks of the LSTM type and Transformer architectures, focused on capturing long-term temporal patterns in the breathing behavior of the user; (ii) the expansion of Respire Bem to other age groups and clinical subgroups; (iii) the longitudinal validation of the efficacy of diaphragmatic breathing and of the predictive architecture in prospective studies; and (iv) the gradual transition from the synthetic database to a prospective clinical population, replacing statistical calibrations with direct clinical and biological measurements, which may support future regulatory evaluations of the system as a decision support tool.

5. Conclusions

This work evaluated an integrated machine learning framework applied to asthma management, articulating two complementary scales of analysis. In Study 1, it was shown that the hybrid architecture composed of ElasticNet, residual CatBoost, direct CatBoost, and adaptive blending, evaluated by walk-forward validation over 30 bimonthly folds, consistently outperformed classical baselines in the daily prediction of respiratory hospitalizations (chapter X of the ICD-10, which includes asthma, COPD, and respiratory tract infections) in the municipality of São Paulo between 2018 and 2022. The model achieved a mean R^2 of 0.675 and a mean skill gain of 41.5% over the seasonal baseline, with robust performance in a time series marked by seasonality, environmental variability, and pandemic impact. The interpretability analysis through SHAP values identified TEMPMED, PM2.5, NO2, and CO as the main predictive factors associated with respiratory hospital admissions, with patterns that can be interpreted in the light of the recent epidemiological literature. The sensitivity analysis confirmed the stability of the findings under variations in the main parametric choices of the pipeline.

 doi:10.20944/preprints202606.1217.v1

In Study 2, the XGBoost and Random Forest models showed the computational feasibility of estimating digital biomarkers related to the expected response to diaphragmatic breathing from 913 qualified records of the Respire Bem system in patients aged 6 to 16 years. XGBoost reached an R^2 of 0.80 for the simulated Asthma Control Test and 0.78 for simulated salivary cortisol, which indicates the ability to capture patterns among adherence, usage behavior, self-reported subjective state, and clinical outcomes generated on the basis of prior evidence. The self-reported sentiment by the patient emerged as the variable with the largest predictive contribution, suggesting that pediatric self-report can be a relevant source of information in digital health systems, especially when combined with objective metrics of duration, frequency, and completeness of the exercises.

Together, the two studies support the feasibility of a dual analytical architecture for asthma management, combining environmental alerts at the population level with remote clinical monitoring at the individual level. Study 1 provides robust predictive validation with real, secondary, and population data, while Study 2 demonstrates an exploratory and methodological stage based on real behavioral data and simulated clinical outcomes. This distinction reinforces the progressive nature of the proposal and points to the need for future prospective validation with direct clinical and biological measurements. The computational reproducibility, the methodological transparency, and the use of official Brazilian sources, such as DATASUS via PCDaS/Fiocruz, CETESB, and INMET, reinforce the potential applicability of the proposed framework to other metropolitan regions of the country, contributing to the advancement of digital health, environmental surveillance, and personalized medicine in the context of respiratory diseases.

Author Contributions: Conceptualization, G.F.S., D.F.A., and D.P.F.; methodology, G.F.S. and D.P.F.; software, G.F.S. and D.P.F.; validation, G.F.S., D.F.A., and D.P.F.; formal analysis, G.F.S. and D.P.F.; investigation, G.F.S. and D.P.F.; resources, D.F.A.; data curation, G.F.S. and D.P.F.; writing, original draft preparation, G.F.S. and D.P.F.; writing, review and editing, G.F.S., D.F.A., and D.P.F.; visualization, G.F.S. and D.P.F.; supervision, D.F.A.; project administration, D.F.A.; funding acquisition, D.F.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Graduate Program in Computational Modeling of the Federal University of Rio Grande. Information on the article processing charge (APC) coverage will be provided to the journal in the appropriate administrative channels upon submission.

Institutional Review Board Statement: Study 1 used publicly available aggregated secondary data and is therefore exempt from review by a Research Ethics Committee, in accordance with Resolutions number 466/2012 and 510/2016 of the Brazilian National Health Council, in particular Article 1, sole paragraph, item III of CNS Resolution 510/2016. Study 2 was approved by the Research Ethics Committee in the Health Area (CEPAS) of FURG (approval number 3,482,646), with registration in the Brazilian Registry of Clinical Trials (ReBEC) under number RBR-3x8py3n.

Informed Consent Statement: Informed consent was obtained from the legal guardians of all pediatric participants involved in Study 2.

Data Availability Statement: The environmental and meteorological data used in Study 1 are publicly available on the CETESB portal (https://cetesb.sp.gov.br) and on the INMET portal (https://portal.inmet.gov.br). The aggregated hospital data were obtained from SIH/SUS through the Data Science Platform Applied to Health of the Oswaldo Cruz Foundation, PCDaS/Fiocruz (https://pcdas.icict.fiocruz.br). The full source code of the pipeline, including the CURATED_CFG dictionary and the per run audit artifacts, is available upon request from the authors. The anonymized data of the Respire Bem system (Study 2) can be obtained from the corresponding authors upon reasonable request and within the applicable ethical restrictions.

Acknowledgments: The authors thank the Graduate Program in Computational Modeling of the Federal University of Rio Grande for the institutional support; CETESB, INMET, and DATASUS for making the data available; the Data Science Platform Applied to Health of the Oswaldo Cruz Foundation (PCDaS/Fiocruz) for the curation of the hospital information; and the patients and guardians who participated in Study 2.

 doi:10.20944/preprints202606.1217.v1

Use of Artificial Intelligence: The authors used a large language model (LLM) assistant exclusively for English-language editing and stylistic refinement of the manuscript drafted in Portuguese. No scientific content, data analysis, modeling decision, result, or conclusion was generated by the AI tool. The authors take full responsibility for the integrity of the work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. World Health Organization. Asthma. Geneva: WHO, 2024. Available online: https://www.who.int/news-room/fact-sheets/detail/asthma (accessed on 4 May 2026).
2. Health Effects Institute. *State of Global Air 2024*; Health Effects Institute, Institute for Health Metrics and Evaluation, UNICEF: Boston, USA, 2024.
3. Moraes, S.L.; Almendra, R.; Santana, P.; Galvani, E. Variáveis meteorológicas e poluição do ar e sua associação com internações respiratórias em crianças: estudo de caso em São Paulo, Brasil. *Cad. Saúde Pública* **2019**, *35*, e00101418. https://doi.org/10.1590/0102-311X00101418. (In Portuguese).
4. Andersen, Z.J.; Vicedo-Cabrera, A.M.; Hoffmann, B.; Melén, E. Climate change and respiratory disease: clinical guidance for healthcare professionals. *Breathe* **2023**, *19*, 220222. https://doi.org/10.1183/20734735.0222-2022.
5. Tran, H.M. et al. The impact of air pollution on respiratory diseases in an era of climate change: a review of the current evidence. *Sci. Total Environ.* **2023**, *898*, 166340. https://doi.org/10.1016/j.scitotenv.2023.166340.
6. Global Initiative for Asthma. *Global Strategy for Asthma Management and Prevention*, 2024 Update; GINA Scientific Committee: Fontana, USA, 2024. Available online: https://ginasthma.org/ (accessed on 4 May 2026).
7. Lee, S.; Ku, H.; Hyun, C.; Lee, M. Machine learning-based analyses of the effects of various types of air pollutants on hospital visits by asthma patients. *Toxics* **2022**, *10*, 644. https://doi.org/10.3390/toxics10110644.
8. Liu, C. et al. Ambient particulate air pollution and daily mortality in 652 cities. *N. Engl. J. Med.* **2019**, *381*, 705–715. https://doi.org/10.1056/NEJMoa1817364.
9. Rackow, B.; König, H.-H.; Wall, M.; Konnopka, C. Co-occurrence of heat and air pollution and their combined effects on health: a systematic review. *Sci. Total Environ.* **2025**, *968*, 180080. https://doi.org/10.1016/j.scitotenv.2025.180080.
10. Souza, A. de; Oliveira-Júnior, J.F. de; Cardoso, K.R.A.; Fernandes, W.A.; Pavao, H.G. The impact of meteorological variables on particulate matter concentrations. *Atmosphere* **2025**, *16*, 875. https://doi.org/10.3390/atmos16070875.
11. Achebak, H.; Garcia-Aymerich, J.; Rey, G.; Chen, Z.; Méndez-Turrubiates, R.F.; Ballester, J. Ambient temperature and seasonal variation in inpatient mortality from respiratory diseases: a retrospective observational study. *Lancet Reg. Health Eur.* **2023**, *35*, 100757. https://doi.org/10.1016/j.lanepe.2023.100757.
12. Romaszko-Wojtowicz, A.; Dragańska, E.; Doboszyńska, A.; Glińska-Lewczuk, K. Impact of seasonal biometeorological conditions and particulate matter on asthma and COPD hospital admissions. *Sci. Rep.* **2025**, *15*, 450. https://doi.org/10.1038/s41598-024-84739-9.
13. Gasparrini, A. Distributed lag non-linear models. *Stat. Med.* **2010**, *29*, 2224–2234. https://doi.org/10.1002/sim.3940.
14. Bochenek, B.; Jankowski, M.; Wieczorek, J.; Gruszczyńska, M.; Sekuła, P.; Jaczewski, A.; Wyszogrodzki, A.; Pinkas, J.; Figurski, M. The impact of ambient air pollution and meteorological factors on emergency hospital admissions of COPD patients in Poland (2012–2019). *Sci. Rep.* **2025**, *15*, 21915. https://doi.org/10.1038/s41598-025-07684-1.
15. Dowlatabadi, Y.; Abadi, S.; Sarkhosh, M.; Mohammadi, M.; Moezzi, S.M.M. Assessing the impact of meteorological factors and air pollution on respiratory disease mortality rates: a random forest model analysis (2017–2021). *Sci. Rep.* **2024**, *14*, 24535. https://doi.org/10.1038/s41598-024-74440-2.
16. Fernandes, S.S. Avaliação do efeito da respiração diafragmática em crianças e adolescentes com asma: ensaio clínico randomizado. PhD thesis, Federal University of Rio Grande, 2023. (In Portuguese).

 doi:10.20944/preprints202606.1217.v1

17. Ferreira, D.P. Respire Bem: uma ferramenta para o tratamento da asma a partir da respiração diafragmática. Master's dissertation, Federal University of Rio Grande, 2022. (In Portuguese).
18. Ferreira, D.P. et al. Use of machine learning algorithms for support and diagnosis of asthma: a systematic literature review. *Rev. Polít. Públicas Cidades* **2026**, *15*, e3055. https://doi.org/10.23900/2359-1552v15n1-45-2026.
19. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, Volume 31; 2018; pp. 6638–6648.
20. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: New York, USA, 2016; pp. 785–794. https://doi.org/10.1145/2939672.2939785.
21. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, Volume 30; 2017; pp. 4765–4774.
22. Tashman, L.J. Out-of-sample tests of forecasting accuracy: an analysis and review. *Int. J. Forecast.* **2000**, *16*, 437–450. https://doi.org/10.1016/S0169-2070(00)00065-0.
23. Bergmeir, C.; Hyndman, R.J.; Koo, B. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Comput. Stat. Data Anal.* **2018**, *120*, 70–83. https://doi.org/10.1016/j.csda.2017.11.003.
24. Cabral-Miranda, W.; Beloni, C.; Lora, F.; Afonso, R.; Araújo, T.; Fernandes, F. Artificial intelligence platform to predict children's hospital care for respiratory disease using clinical, pollution, and climatic factors. *J. Glob. Health* **2025**, *15*, 04207. https://doi.org/10.7189/jogh.15.04207.
25. Barnett-Itzhaki, Z.; Nir, V.; Kellner, A.; Biton, O.; Toledano, S.; Klein, A. Machine learning models for predicting pediatric hospitalizations due to air pollution and humidity: a retrospective study. *Pediatr. Pulmonol.* **2025**, *60*, e71106. https://doi.org/10.1002/ppul.71106.
26. Monteiro Martins, L.; Coz, E.; Maucort-Boulch, D.; Hacid, M.-S. Machine learning with environmental predictors to forecast hospital visits and admissions: a systematic review. *Environ. Syst. Res.* **2025**, *14*, 28. https://doi.org/10.1186/s40068-025-00401-x.
27. Bhaskaran, K. et al. Time series regression studies in environmental epidemiology. *Int. J. Epidemiol.* **2013**, *42*, 1187–1195. https://doi.org/10.1093/ije/dyt092.
28. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, USA, 2009.
29. Hyndman, R.J.; Athanasopoulos, G. *Forecasting: Principles and Practice*, 2nd ed.; OTexts: Melbourne, Australia, 2018.
30. Hodson, T.O. Root mean square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geosci. Model Dev.* **2022**, *15*, 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022.
31. Hyndman, R.J.; Koehler, A.B. Another look at measures of forecast accuracy. *Int. J. Forecast.* **2006**, *22*, 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001.
32. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. https://doi.org/10.1038/s42256-019-0138-9.
33. Donoho, D. 50 years of Data Science. *J. Comput. Graph. Stat.* **2017**, *26*, 745–766. https://doi.org/10.1080/10618600.2017.1384734.
34. Wang, W.; Luo, X.; Ren, Z.; Fu, X.; Chen, Y.; Wang, W.; Bao, Y.; Zheng, Y.; Cao, K.; Chen, J. Impact of COVID-19 pandemic measures on hospitalizations and epidemiological patterns of twelve respiratory pathogens in children with acute respiratory infections in southern China. *BMC Infect. Dis.* **2025**, *25*, 103. https://doi.org/10.1186/s12879-025-10463-y.
35. Hyrkas, H.; Ikäheimo, T.M.; Jaakkola, J.J.K.; Jaakkola, M.S. Asthma control and cold weather-related respiratory symptoms. *Respir. Med.* **2016**, *113*, 1–7. https://doi.org/10.1016/j.rmed.2016.02.005.
36. Li, X. et al. Lag effect of ambient temperature on respiratory emergency department visits in Beijing: a time series and pooled analysis. *BMC Public Health* **2024**, *24*, 1363. https://doi.org/10.1186/s12889-024-18839-6.

 doi:10.20944/preprints202606.1217.v1

37. Chen, J.M.; Zovko, M.; Šimurina, N.; Zovko, V. Fear in a handful of dust: the epidemiological, environmental, and economic drivers of death by PM2.5 pollution. *Int. J. Environ. Res. Public Health* **2021**, *18*, 8688. https://doi.org/10.3390/ijerph18168688.
38. Putaud, J.-P. et al. Impact of 2020 COVID-19 lockdowns on particulate air pollution across Europe. *Atmos. Chem. Phys.* **2023**, *23*, 10145–10161. https://doi.org/10.5194/acp-23-10145-2023.